DOCTOR OF MEDICINE

Genetic susceptibility to rheumatoid arthritis

MacKay, Kirsten Robyn

Award date:
2003

Awarding institution:
University of Bath

Genetic susceptibility to rheumatoid arthritis

Submitted by Kirsten Robyn MacKay MB ChB, MRCP

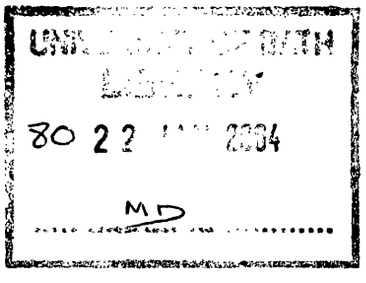
for the degree of M.D.

of the University of Bath

2003

COPYRIGHT

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346


80 2 2
MD

2 February 2002

BMJ

Rheumatoid arthritis: is it genetic?

Probably not p264

This was the front cover of the British Medical Journal (BMJ) in February 2002. The following thesis set out to clarify this question

Contents

Acknowledgements

Many people have helped me to complete this project all of who deserve my sincerest thanks. In particular, I would like to thank my supervisor at the Wellcome Trust Centre for Human Genetics and the Nuffield Department of Medicine, Professor Paul Wordsworth, for his valuable help and insight. My other supervisor, Dr Andrei Calin from the Royal National Hospital for Rheumatic Diseases and University of Bath has helped over many years with advice and encouragement enabling me to lay the groundwork for this project. I must also thank my generous sponsor, the Arthritis Research Campaign, for giving me the opportunity to undertake this M.D.

I have collaborated with a number of people whilst undertaking this project and in particular those working at the ARC Epidemiology Unit based at Manchester University deserve my thanks. They include Professor William Ollier, Drs Jane Worthington and Anne Barton, Anne Myerscough and Stephen Eyre. I would also like to thank the many people at the Wellcome Trust for Human Genetics who helped in so many ways during my three years in Oxford. Special thanks go to Dr Stephen Laval for his patience, Dr Matthew Brown for his statistical help, Anita Milicic and Dorothea Lee for their technical assistance and members of the bio-informatics department who helped me sort out my frequent IT problems and understand the vagrancies of the statistical packages.

Finally I must thank my friends and family without whom I would not have completed this task. Particular thanks must go to Drs Julia Newton, Suzanne Donnelly and Jon Packham for their support during my time in Oxford and to my new Portsmouth friends and colleagues for their support whilst I have been writing up. Most especially, thank you to my parents who have loved, encouraged and supported me throughout my life and sincere thanks to Dr Anthony Collins for his help, encouragement, support and friendship over the last five years. I hope I have managed to make you all proud of me, I could not have done it without you.

Collaboration

Some of the work contained within this thesis was undertaken in collaboration with the Arthritis Research Campaign Epidemiology Unit (ARC-EU), University of Manchester. The sibling pairs affected by rheumatoid arthritis used for the two studies involving linkage analysis were identified from the Arthritis and Rheumatism Campaign United Kingdom National Repository of family material. Genotyping of some microsatellite markers (on chromosomes 8 – 15, 17 – 20 and X chromosome) was performed by Stephen Eyre, Anne Myerscough and Dr. Anne Barton (ARC-EU, University of Manchester).

The South African patients included in the Interleukin 10 case control study were recruited by Dr M. Tikly, (Chris Hani Baragwanath Hospital and University of the Witwatersrand, PO Bertsham 2013, South Africa). The United Kingdom patients used in the case control study were recruited by Professors R. Sturrock, D.G.I. Scott, P. Bacon and Dr R Madock. Singleton RA families were identified from the Arthritis and Rheumatism Campaign United Kingdom National Repository of family material or from an Oxford based recruitment drive undertaken by Mrs. Helen Yates. HLA typing was performed on Oxford patient samples by Dr J. Newton.

I was assisted in some of the laboratory work by Anita Milicic and Dorothea Lee (PCR and gel electrophoresis) but all project management, collaboration and statistical analysis was undertaken by myself.

List of plates

List of figures

List of tables

Glossary of terms

ABI	Applied Biosystems
ACR	American College of Rheumatology
ARA	American Rheumatism Association
ARC	Arthritis Research Campaign
ARC-ERU	Arthritis Research Campaign – Epidemiology Research Unit
AS	ankylosing spondylitis
ASP	affected sibling pair
cM	centiMorgan
Con A	concanavalin A
CRH	corticotrophin releasing hormone
DMARD	disease modifying anti-rheumatic drug
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DZ	diazygotic
ECRAF	European Consortium of RA Families
ER	endoplasmic reticulum
ESR	erythrocyte sedimentation rate
FDI	functional disability index
FGF	fibroblast growth factor
GI	gastrointestinal
GNP	gross national product
HAQ	health assessment questionnaire
hsp	heat shock protein
HLA	human leucocyte antigen
ibs	identity by state
ibd	identity by descent
IDDM	insulin-dependent diabetes mellitis
IFN	interferon
IL	interleukin

IL-1ra	interleukin receptor antagonist
JIA	juvenille idiopathic arthritis
LOD	log of the odds for linkage
LPS	lipopolysaccharide stimulation
M	Morgan
Mb	megabase
MCP	metacarpophalangeal joints
Mg^{2+}	magnesium
$MgCl_2$	magnesium chloride
MHC	major histocompatibility complex
MLS	maximum LOD score
mRNA	messenger ribonucleic acid
MZ	monozygotic
NARAC	North American RA Consortium
NOAR	Norfolk Arthritis Register
NPL	non-parametric linkage
OA	osteoarthritis
OCP	oral contraceptive pill
PBMC	peripheral blood mononuclear culture
PCR	polymerase chain reaction
PDGF	platelet derived growth factor
PG	prostaglandin
PIP	proximal interphalangeal joints
QTL	quantitative trait locus
RA	rheumatoid arthritis
RCGP	Royal College of General Practitioners
RF	rheumatoid factors
RFLP	restriction length polymorphisms
SE	shared epitope
SLE	systemic lupus erythematosus
SNP	single nucleotide polymorphism

SMR	standardised mortality ratio
Taq	*Thermus aquaticus* DNA polymerase
TCR	T cell receptor
TDT	transmission disequilibrium testing
TIMP	tissue inhibitor of metalloproteinase
TNFα	tumour necrosis factor α
UK	United Kingdom
US	United States
WGS	whole genome screen
WTCHG	Wellcome Trust Centre for Human Genetics
χ^2	chi-squared
λ_s	Sibling recurrence risk
θ	recombination fraction

Abstract

The heritability of Rheumatoid Arthritis is approximately 60%. Although association with the HLA region is well recognised, these genes account for only 40% of the total genetic contribution. Linkage mapping and association studies need to proceed in parallel to identify which non-MHC genes are involved in the remaining 60% of the genetic contribution. The work described in this thesis represents a systematic approach to identifying these non-HLA effects, using genome-wide linkage mapping and candidate gene methods.

To identify regions exhibiting genetic linkage to rheumatoid arthritis a systematic, whole genome linkage analysis was undertaken. Two hundred and fifty-one affected sibling pairs from 182 United Kingdom families were studied using 365 highly informative microsatellite markers. Highly significant linkage was identified around the HLA region on chromosome 6 (max LOD = 4.8 at 44.9cM, p=0.000001). Eighteen other sites of nominal linkage ($p<0.05$) were identified on chromosomes 1, 2, 3, 4, 6, 7, 10, 12, 14, 16, 21 and the X chromosome by single point analysis (23 markers). Eight of the non-MHC regions (on chromosomes 1, 6, 7, 14, 16, 21, and X) also showed evidence of linkage by multi-point analysis.

A parallel linkage study designed to replicate the 25 regions of nominal linkage ($p<0.05$) reported following a genome-wide linkage study of 97 European affected sibling pair families was also undertaken. Fifty-nine microsatellites within the 25 regions of interest (including IDDM6 on chromosome 18 and IDDM9 on chromosome 3) were used to genotype 368 affected sibling pairs from 280 families. Markers on chromosomes 12, 15, and 21 (d12s95, CYP 19, d21s1252) showed evidence of nominal linkage with p values ≤ 0.05. Markers close to IDDM 6 on chromosome 18 showed p values of between 0.1 and 0.5, not lending additional support to a locus near IDDM6 in RA.

Interleukin 10 (IL10), an immunoregulatory cytokine, is a potent up-regulator of B cell production and differentiation but has anti-inflammatory capabilities and can directly down-regulate TNF, IL1, IL8 and IFNγ production. Data from twin and family studies suggest large inter-individual variations in secretion which are 75% heritable and as such IL10 is a plausible candidate gene for involvement in RA. It is highly polymorphic with point mutations in the promoter region and two microsatellite loci IL10.R and IL10.G, 1.1 and 4kb upstream of the transcription initiation site. Higher levels of IL10 secretion have been associated with allele 2 of IL10.R (IL10.R2) and IL10.R3 has been associated with decreased secretion. Additionally, a case-control study including two independent Caucasian populations and one African-American cohort found an over-representation of the IL10.R2 allele with a concomitant reduction of IL10.R3 in all three rheumatoid arthritis populations.

Three studies investigating IL10 were undertaken. Two case-control studies including two cohorts of racially distinct RA patients (186 UK Caucasians with severe rheumatoid arthritis and 138 South Africans of Zulu or Sotho origin) were performed. An association with RA was not confirmed in either study but demonstrated significantly different frequencies of the IL10.R2 allele in the two study populations. The third study was a family-based association study and included 163 probands and their families. Single marker and haplotypic association analysis was performed by transmission disequilibrium testing (TDT analysis) using the software package TRANSMIT. The IL10.R1 allele was transmitted to affected individuals more frequently than expected ($p<0.05$) and the IL10.R3 allele was transmitted less frequently than expected ($p<0.05$). This effect was particularly pronounced with the IL10.R3/IL10.G10 haplotype ($p<0.005$). The still stronger negative association identified with the IL10.R3/IL10.G10 haplotype suggested it was not IL10.R itself but another polymorphism on the particular haplotype that may be primarily involved with RA.

The application of these methods to the examination of the genetic component of RA is discussed. Plans for future work include systematic genome-wide screening of positional candidates and evaluation of candidates from studies of other human and animal models of inflammatory disease.

Chapter 1 – section 1

Plate A is a photograph of an 18th century oil painting by William Hoare depicting a rheumatology and orthopaedic combined clinic at the Mineral Water hospital in Bath. Dr William Olliver and Mr Jeremiah Peirce are reviewing a child with a skin disease, a man with a wrist drop and woman with inflammatory arthritis. Between 1752 - 1756 one third of patients seen at the Bath hospital had musculoskeletal conditions. In 1935 the hospital was renamed the Royal National Hospital for Rheumatic diseases as by then all patients were referred with musculoskeletal diseases.

CHAPTER 1 - section 1

1.1.1 Rheumatoid Arthritis: a natural history.

Hippocrates first described rheumatic diseases in the fourth century B.C. The term 'rheuma' was used in the first century A.D. to describe a flow of pain through the joints of the body (Sangha, 2000). Although the study of ancient North American skeletons suggest that rheumatoid arthritis (RA) existed at least 3000 years ago (Goemaere *et al.*, 1990; Sangha, 2000), the first recorded clinical description of RA is normally credited to Augustin-Jacob Landré-Beauvais in his thesis written in 1800. Archibald Garrod coined the term 'rheumatoid arthritis' (RA) in 1859 but included inflammatory polyarthritis and polyarticular osteoarthritis (Garrod, 1859; Halberg, 1998). By 1922 polyarticular osteoarthritis was excluded from the definition (Parish, 1963) as were the sero-negative arthritides in 1958 (Ropes *et al.*, 1958).

Classification of the rheumatic diseases has been hampered by the absence of firm evidence for disease aetiology. With no unique clinical or laboratory features, disease classification has been dependent upon a combination of common clinical, radiological and laboratory findings. The use of predominately clinical criteria to define these diseases has inevitably resulted in broad categories of differing rheumatic conditions with considerable clinical and pathological overlap. Diagnosis of a particular condition can therefore be difficult in the early stages of disease (Sangha, 2000). For example, it has been reported that RA was not diagnosed at presentation in up to twenty per cent of patients with polyarthritis attending an early synovitis clinic (Salmon *et al.*, 1993; Wordsworth and Brown, 1997).

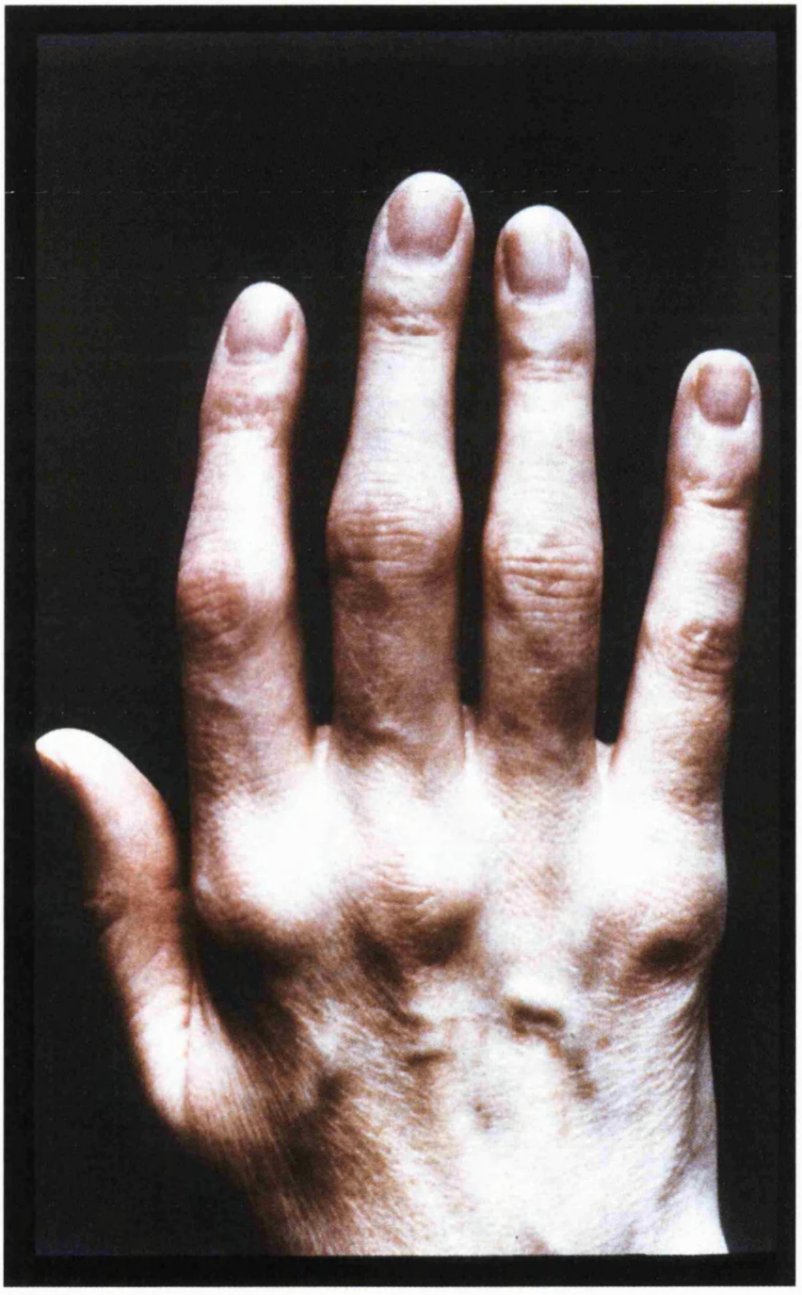


Plate 1.1.1 shows a hand with MCP and PIP joint swelling. This hand shows the changes of relatively recent onset RA.

1.1.1.1 Rheumatoid arthritis: clinical features.

RA is a chronic systemic inflammatory condition predominately involving synovial joints and affecting up to 1% of the population. The cause is unknown and the disease is clinically heterogenous with a wide range of disease severity. Most people affected (~80%) have elevated titres of serum rheumatoid factors. The majority develop a chronic form of the disease characterised by relentless progressive joint destruction and disordered immune function leading to marked functional impairment and disability. Some people experience only mild oligoarticular synovitis and the disease resolves spontaneously within two to three years of onset. Yet in others extra-articular features are prominent (Matteson *et al.*, 1998).

1.1.1.2 Rheumatoid arthritis: presentation

Disease onset and the subsequent clinical course vary considerably, hence the measurement of disease activity or prognosis is not easily defined. A gradual onset is most common (at least 50%) whilst a minority (10 – 25%) developing an acute widespread symmetrical polyarthropathy over a few hours or days (Gordon and Hastings, 1998). The classical presentation of RA initially involves a distal symmetrical polyarthritis affecting the small joints of the hands and feet with associated fatigue and early morning stiffness [Gordon, 1998] (plates 1.1.1, 1.1.2 and 1.1.3). Bone erosions, leading to joint destruction generally develop within three years of disease onset [van der Heijde *et al.*, 1992] (plates 1.1.5 and 1.1.6). Frequently, the arthritis spreads to involve proximal joints and can lead to severe functional disability if weight-bearing joints are affected (Gordon and Hastings, 1998). Many require joint replacements as a consequence of joint damage resulting from inadequately controlled disease (Gordon and Hastings, 1998).

RA is a multi-system disease and constitutional features, such as fatigue and weight loss, may occur early in the course of the disease and may

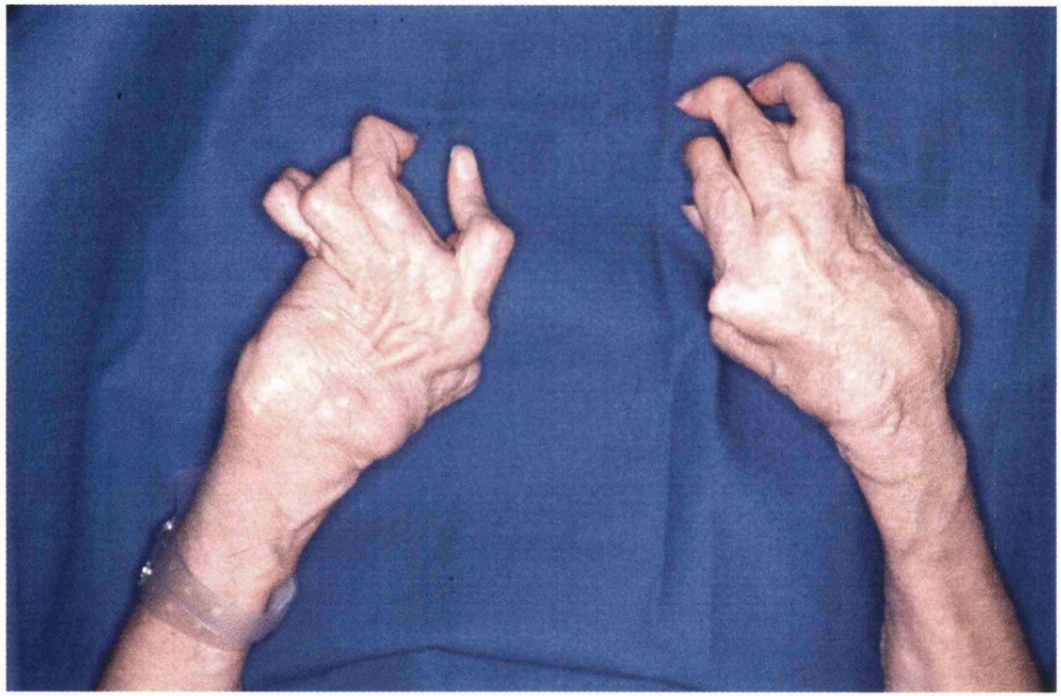


Plate 1.1.2 shows the hands and wrist changes that can occur in severe RA. Subluxation and ulnar deviation can be seen at the wrist and MCP joints. Boutonnieres deformity and swan necking can be seen in the fingers.

predominate over shadowing the joint manifestations (Matteson *et al.*, 1998). Associated non-articular features include subcutaneous nodules (~20%) (Matteson *et al.*, 1998) (plate 1.1.4), vasculitis (plate 1.1.7), pericarditis, pulmonary nodules or interstitial fibrosis, normochromic normocytic anaemia, thrombocytosis, episcleritis or scleritis. Between 10 – 35% of patients develop secondary Sjögren's syndrome with conjunctivitis sicca and xerostomia (Matteson *et al.*, 1998). Pleurisy and pericarditis are common. Up to 50% of patients have echocardiographic evidence of pericarditis but most have no overt symptoms (Matteson *et al.*, 1998). Rheumatoid vasculitis may result in extensive cutaneuous ulceration, a peripheral neuropathy (with a glove and stocking sensory loss), mononeuritis multiplex and myocardial or bowel ischaemia. As many as one in nine men with RA sustain an episode of major vasculitis during their lifetime (Wordsworth and Brown, 1998). Less than 1% of patients develop Felty's syndrome, a combination of rheumatoid disease, neutropenia, splenomegaly and lymphadenopathy (Matteson *et al.*, 1998). This rarely develops within ten years of disease onset and is frequently associated with pronounced extra-articular features. Overwhelming infection reduces the life expectancy in this group.

1.1.2 Epidemiology of rheumatoid arthritis

Epidemiological studies of RA are dependent on the criteria used to define the disease (Sangha, 2000). However, choosing the criteria used for diagnosis is difficult without the identification of an aetiological agent or specific clinical or laboratory feature to define the disease. Hence, various different criteria have been produced in an effort to improve diagnostic sensitivity and specificity. Therefore, the criteria used to diagnose an individual with RA have changed over the years making it difficult to compare disease prevalence over the decades and hampering most population

Classification of Rheumatoid Arthritis - The American Rheumatism Association 1958 diagnostic criteria

Categories	Criterion	Comment
Possible rheumatoid arthritis	Morning stiffness	Duration: ≥1 hr and lasting ≥ 6 weeks
Probable rheumatoid arthritis	Arthritis in at least 3 joint areas *	With soft tissue swelling or fluid - lasting ≥ 6 weeks
Definite rheumatoid arthritis	Arthritis of hand joints	Wrist, MCP +/or PIP joints - lasting ≥ 6 weeks
Classical rheumatoid arthritis	Symmetrical arthritis *	At least one area - lasting ≥ 6 weeks
5	Subcutaneous rheumatoid nodules	Observed by a physician
6	Radiographic changes typical of RA	Seen on an anteroposterior radiograph of the hands and wrists
7	Serum rheumatoid factor	Assessed by a method positive in < 5% of control subjects

*includes right or left proximal interphalangeal (PIP) joints , metacarpophalangeal (MCP) joints, metatarsophalangeal (MTP) joints, wrist joint, elbow joint, knee joint or ankle joint

Table 1.1.1. Adapted from Ropes et al [Ropes, 1958 ›] summarises the classification criteria for RA published by the American Rheumatism Association (ARA) in 1958. These were defined in 1956 following the analysis of 332 cases and 11 criteria with 19 exclusions were proposed. The criteria were revised in 1958. "Classic" RA required at least 7 criteria, whereas "definite" RA required at least 5 criteria and 6 weeks of joint symptoms. "Probable" RA required at least 3 criteria and 6 weeks of joint symptoms.

studies. Until recently, the most widely used diagnostic criteria were published by the American Rheumatoid Association (ARA) in 1958 (Ropes *et al.*, 1958). These criteria defined RA with varying degrees of certainty in terms of "classic", "definite", "probable" and "possible" disease (table 1.1.1). This complexity was simplified in the American College of Rheumatology (ACR) 1987 revised criteria for the classification of RA (Arnett *et al.*, 1988) when at least four criteria had to be fulfilled for the diagnosis of RA to be made (table 1.1.2). The criteria distinguished RA from other rheumatic diseases (such as systemic lupus erythematosus) with a specificity ranging from 85 – 95% and sensitivity between 77 – 95% (MacGregor *et al.,* 1998; Arnett *et al.*, 1988). It is important to note that the 'gold standard' used to test these criteria against and so determine the sensitivity and specificity was a diagnosis made by a physician. Also, as they were defined in a secondary care setting (where individuals studied definitely had RA or another defined musculoskeletal condition) they do not perform well in diagnosing RA in an early arthritis cohort where initial symptoms and signs are not so well defined. In a community-based setting (NOAR) they had little more than a random ability to predict persistence of arthritis, the development of radiological erosions or moderate disability (Harrison et al, 1998). The 1987 ACR criteria were modified for epidemiological study by accepting 'deformity' in lieu of 'current swelling' as one of the 4/7 criteria needed (table 1.1.3). This modification improved sensitivity and specificity in these circumstances (section 1.1.3) (MacGregor *et al.*, 1994a).

Disease occurrence is assessed by two measures, both of which present methodological problems for those studying RA. Prevalence refers to the number of existing cases, and incidence, the rate of new cases arising in a given period (Silman and Hochberg, 1993). Studies of incidence require large samples and prolonged follow-up as incidence rates are fairly low in RA. As yet, no criteria specific for early RA have been described thus a number of early cases may be missed. Ideally studies of prevalence should include all

The American College of Rheumatology 1987 revised diagnostic criteria for the classification of Rheumatoid Arthritis

	Criterion	Comment
1	Morning stiffness	Duration: ≥1 hr and lasting ≥ 6 weeks
2	Arthritis in at least 3 joint areas *	With soft tissue swelling or fluid - lasting ≥ 6 weeks
3	Arthritis of hand joints	Wrist, MCP +/or PIP joints - lasting ≥ 6 weeks
4	Symmetrical arthritis *	At least one area - lasting ≥ 6 weeks
5	Subcutaneous rheumatoid nodules	Observed by a physician
6	Serum rheumatoid factor	Assessed by a method positive in < 5% of control subjects
7	Radiographic changes typical of rheumatoid arthritis ^	Seen on an anteroposterior radiograph of the hands and wrists

* (criteria 2 and 4) includes right or left proximal inter-phalangeal (PIP) joints, metacarpo-phalangeal (MCP) joints, wrist, elbow, knee, ankle or metatarso-phalangeal (MTP) joints.
^ (criteria 7) must include erosions or unequivocal bony decalcification (peri-articular osteoporosis) localised to or most marked adjacent to the involved joints

Table 1.1.2. Adapted from Arnett et al (Arnett *et al*, 1988) summarises the American Rheumatism Association (ARA) 1987 revised criteria for the classification of RA. The criteria were defined following the comparison of 262 individuals with RA and 262 control subjects. The sensitivity and specificity of the criteria for RA were 91-94% and 89% respectively. The categories are different from the 1958 ARA criteria in that RA is defined by the presence of 4 or more criteria, and no further qualifications (classic, definite, or probable) or list of exclusions are required.

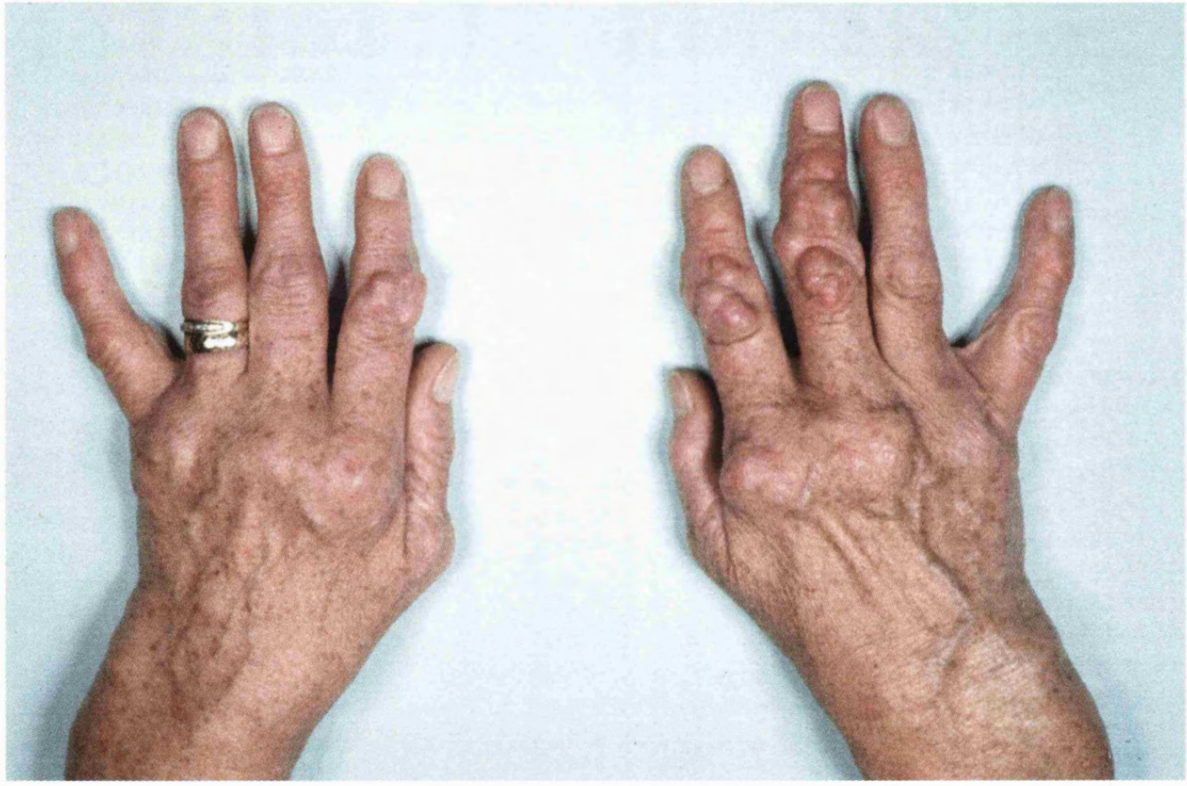


Plate 1.1.4 shows the hands of an individual with RA and rheumatoid nodules.

past and inactive cases of RA. However, to achieve this the use of criteria specifically designed to recognise quiescent disease are required (table 5.1.3). It is frequently difficult to establish the extent to which remitted disease has been included in published reports (MacGregor *et al.,* 1998).

1.1.2.1 Rheumatoid arthritis: prevalence

RA occurs throughout the world. Most studies in Caucasian, European and North American populations report a point prevalence of between 0.5 and 2% (table 1.1.5) (Spector *et al.,* 1990; MacGregor *et al.,* 1998). Prevalence rates have been based on large cross-sectional population samples and age-specific prevalence increases in all studies (MacGregor and Silman, 1998) (Symmons *et al*, 2002). Prevalence rates are higher in women and it has been estimated that the overall male to female ratio of affected individuals is ~3:1. However, at age 30 years the ratio is 6:1 but this falls almost to equality by age 60 years (Wordsworth and Brown, 1997).

1.1.2.2 Rheumatoid arthritis: incidence

Few incidence studies have been reported because of the difficulty in establishing an early diagnosis. Most studies have been retrospective, although there have been two recent prospective studies of incidence rates in RA (table 1.1.6). The Norfolk Arthritis Register has been able to derive incidence rates from a population registry of all individuals with polyarthritis presenting to hospital and general practitioners in Norfolk (Symmons *et al.*, 1994), (Wiles *et al*, 1999). Dugowson et al identified new cases occurring in women attending a health maintenance program in Seattle, USA between 1987 and 1989 (Dugowson *et al.*, 1991). The annual overall incidence of RA ranges from 0.24 to 0.29/1000. For males it is between 0.14 to 0.2/1000 and for females from 0.36 to 0.5/1000 (Dugowson *et al.*, 1991; Linos *et al.*, 1980; MacGregor and Silman, 1998; Symmons *et al.*, 1994). RA may present at almost any time of life but the peak incidence is between the ages of 45 – 75 years (Symmons *et al.*, 1994) (Wiles *et al*, 1999).

Incidence of Rheumatoid Arthritis

Geographical area	Study design	Year	Incidence/1000 person years		
			Combined	Females	Males
Sudbury, Massachusetts USA [O'Sullivan, 1968	New cases between 2 cross-sectional surveys	1964 - 1965	-	0.29	-
Olmsted County, USA [Linos, 1980	Retrospective – hospital attenders	1950 - 1974	0.37	0.48	0.22
RCGP, UK (1979)	Prospective – new episodes in primary care	1970 - 1972	2.99	4.20	1.63
RCGP, UK (1986)	Prospective – new episodes in primary care	1981- 1982	2.47	3.34	1.50
Seattle, Washington (Dugowson, 1991)	Prospective – new referrals*	1987 - 1989	-	0.24	
Norfolk, UK (Symmons, 1994)	Prospective – community ascertained cases §	1992		0.36	0.15
Norfolk, UK (Wiles, 1999)	Prospective – community ascertained cases (5 year data) §	1999		0.54	0.24

RCGP = Royal College of General Practictioners
* = new female RA cases attending a health maintenance organisation
§ = all cases of inflammatory polyarthritis presenting to general or hospital practitioners

Table 1.2.6 adapted from MacGregor and Silman (MacGregor, 1998) summarises the available data used to assess the incidence of RA.

1.1.2.3 Rheumatoid arthritis: geographical variation

Although similar estimates of the prevalence of RA have been obtained from diverse populations worldwide there are exceptions (table 1.1.5). Despite fairly standard levels of RA in the urban black South African population the disease appears to be exceptionally rare in rural Sub-Saharan Africa. It is rare in both urban and rural Chinese populations. However, certain native North American Indians including the Pima, Chippewa and Yakima have a much higher prevalence at ~5% although the disease is rarely seen in other groups such as the Blackfeet or Haida Indians (MacGregor *et al.,* 1998; Silman, *et al.* 1993) (table 1.1.5). The relative genetic factors influencing this geographical variation are discussed later (section 1.2.3) but it is likely that environmental factors are also involved. An association with urban industrial environments has been suggested by the low disease prevalence rate in rural South African blacks when compared with the higher frequency of disease in their urban counterparts (Beighton *et al.*, 1975; Solomon *et al.*, 1975), but this has not been confirmed in other populations. For instance urban Hong Kong Chinese have the same low prevalence as those in rural Kinmet (Silman and Hochberg, 1993) and blacks living in inner-city Manchester have a lower prevalence than whites living in the same area (MacGregor *et al.*, 1994b). It is possible that the rural environment could be protective and it has been suggested that chronic stimulation of the immune system by tropical infections including malaria may in some way be protective against the development of autoimmunity (Greenwood, 1968).

1.1.2.4 Rheumatoid arthritis: variation over time

In some populations the incidence of RA seems to be declining. For instance, the Rochester Epidemiology Program data show a decline in female incidence from 1960 onwards (Linos *et al.*, 1980) and the incidence in Pima Indians has halved in both males and females between 1965 and 1990 (Jacobsson *et al.*, 1994). However, data from a Finnish population register

Prevalence of definite Rheumatoid Arthritis in European, North American, African and Asian populations

Continent	Population	Prevalence (%)
Europe	Finland	2.0
	England	1.1
	Netherlands	0.9
	Denmark	0.8
	Bulgaria	0.8
North America	US National Health Survey	1.0
Native Populations	Chippewa Indians, USA	5.3
	Inuit, Canada	0.6
	Inpiat Eskimos, USA	1.0
	Pima Indians, USA	5.3
Africa	Liberia and Nigeria	0.1
	South Africa (urban)	0.9
	South Africa (rural)	0.1
Asia	China	0.3
	Japan	0.6

Table 1.1.5 adapted from Wordsworth and Brown (Wordsworth, 1997) lists the differing estimates of prevalence of definite RA in various adult populations including European, North American, African and Asian populations. It can be seen from this table that the prevalence of RA varies from 0.1% in the rural South African population to over 5% in some North American Indian populations. Of note, the prevalence of RA in the South African urban population is nearly 1% which is in the same order of magnitude to the prevalence reported by the US National Health Survey and throughout most European populations. This figure is at odds with that of the South African rural population.

showed the incidence remained static between the years 1970 to 1980 (MacGregor, *et al.* 1998). The severity of RA may also be declining over time. Silman et al analysed successive birth cohorts and found peaks in erosive, seropositive and nodular disease in individual presenting with RA in the 1960s. Subsequently, the severity of the condition has been in decline (Silman *et al.*, 1983).

1.1.2.5 Rheumatoid arthritis: risk factors

a) Host factors

Given the predominance of rheumatoid arthritis in women many studies have investigated the association of sex hormones, menstrual and reproductive factors in relation to disease onset, severity and remission. It is well established that pregnancy is associated with disease remission and exacerbations are common in the postpartum period (Nelson *et al.*, 1993; Persellin, 1976). Several studies have suggested that nulliparity is a risk factor (Spector *et al.*, 1990) although no association was found in a recent large prospective cohort study from Finland (Heliovaara *et al.*, 1995). In a prospective study of 23,000 United Kingdom women the oral contraceptive pill (OCP) was associated with halving the incidence of rheumatoid arthritis (Wingrave and Kay, 1978) but further studies have been unable to confirm this. A meta-analysis concluded that although OCP use may not protect against disease it may postpone disease onset (Spector and Hochberg, 1990). One large study investigating age of disease onset found that the average women developed the first symptoms of RA at the time of her menopause (Goemaere *et al.*, 1990).

b) External environment

There is no convincing epidemiological data to support the role of infection. Cases of RA do not cluster over space or time and concordant monozygotic twins, sibling pairs and spouse couples have not shown any similarity in the

Plate 1.1.5 is a radiograph of the hands and wrists of an individual with RA. Erosions can be seen particularly in the wrists.

timing of disease onset. However, it is conceivable that a ubiquitous infectious agent may be responsible for RA in a genetically susceptible host. Animal models have suggested a number of plausible mechanisms. Specific agents have been isolated from human synovium and homology exists between the antigenic components of infectious organisms, synovium and cartilage (Phillips, 1988). Raised titres of Epstein-Barr virus have been demonstrated in those with RA, but in one small longitudinal study titres were raised prior to disease onset (Kouri *et al.*, 1990). Raised titres of parvovirus and cytomegalovirus have also been reported but not consistently in all studies (Hajeer *et al.*, 1994; Walker *et al.*, 1987). No conclusive evidence of association has been found with numerous other putative infectious agents such as mycoplasma, proteus, rubella or mycobacteria (Walker *et al.*, 1987).

Some studies have associated low socio-economic status with a worse disease outcome but conflicting data are available regarding the influence on disease susceptibility. Recently, prospective data from the Norfolk Arthritis Register have shown no association with social class or economic status (Bankhead *et al.*, 1996). Specific occupations such as coal mining, granite working and those handling organic solvents have been implicated but these reports have never been confirmed (MacGregor *et al.*, 1998; Silman *et al.*, 1993).

1.1.3 Classification criteria modified for genetic and epidemiological studies

In genetic or epidemiological studies the failure to recognise "cases" with inactive or mild disease has lead to the misclassification of disease status and has important consequences. In family studies it may lead to an underestimation of disease recurrence risks. The presence or absence of

disease needs to be ascertained through a single encounter with a study subject. Hence, the classification of disease status needs to be consistent, incorporating disease features present during current examination but also those that have occurred in the past (e.g. the difference between current synovitis and evidence of past synovitis). The information regarding past disease activity can be given by the patient or come from clinical records. Many existing criteria have shortcomings as they do not specifically take into account features of inactive disease. The 1958 ARA and 1987 ACR diagnostic criteria (Arnett *et al.*, 1988; Ropes *et al.*, 1958) were developed to distinguish RA from other rheumatic conditions and are weighted towards the recognition of currently active disease (tables 1.1.1 and 1.1.2). These diagnostic criteria are considered to have high levels of sensitivity and specificity but were initially tested in a hospital setting where the proportion of active disease is inevitably higher than in the community where there may be a relatively large number of individuals with inactive disease. The lack of a specific diagnostic test for RA also presents problems as the study population, unless carefully selected, may exhibit considerable clinical heterogeneity. MacGregor and colleagues compared seven different sets of RA criteria and compared them to the 'gold standard' of a physician diagnosis. The individuals included were participating in an epidemiological study of twins, where at least one of the pair had RA (MacGregor *et al.*, 1994a). Differences in disease classification substantially altered the estimation of disease concordance for RA in the twin pairs studied. Only 124 of the 283 individuals reporting polyarthritis fulfilled all seven sets of disease criteria. Differences were most obvious between the criteria that accepted 'current swelling' only and those accepting 'current swelling' or 'deformity' as part of the criteria set. The 'current swelling only' criteria recorded point prevalence or currently active disease whereas the 'current swelling or deformity' criteria recorded cumulative prevalence or "ever disease". When "ever disease" was taken into account sensitivity and specificity increased. Overall, the 1987 ACR criteria modified for epidemiological study performed

The American College of Rheumatology 1987 revised criteria for the classification of Rheumatoid Arthritis – modified for population and family studies

	Criterion	Comment
1	Morning stiffness	Duration: ≥1 hr - at any time during the disease course
2	Arthritis in at least 3 joint areas *	With soft tissue swelling or fluid - currently or documented previously
3	Arthritis of hand joints	Swelling of wrist, MCP or PIP joints - currently or documented previously
4	Symmetrical arthritis *	Simultaneous bilateral involvement of the same area with swelling - currently or documented previously
5	Subcutaneous rheumatoid nodules	Observed by a physician - currently or documented previously
6	Serum rheumatoid factor	Assessed by a method positive in < 5% of control subjects - currently or documented previously
7	Radiographic changes §	Seen on an anteroposterior radiograph of the hands and wrists

* (criteria 2 and 4) includes right or left proximal inter-phalangeal (PIP) joints, metacarpo-phalangeal (MCP) joints, wrist, elbow, knee, ankle or metatarso-phalangeal (MTP) joints.
^ (criteria 7) must include erosions or unequivocal bony decalcification (peri-articular osteoporosis) localised to or most marked adjacent to the involved joints

Table 1.1.3. Adapted from MacGregor et al (MacGregor *et al*, 1995) summarises the American Rheumatism Association (ARA) 1987 revised criteria for the classification of RA modified for population and family studies. These criteria are very similar to the original 1987 revised criteria (table 1.1.2) but vary in that previously documented arthritis, subcutaneous nodules or positive rheumatoid factor may be included in the criteria even if the individual is currently in remission (i.e. without current soft tissue swelling or joint fluid). Hence these criteria identify both individuals who have currently active disease and those who have had active RA in the past.

best in terms of sensitivity and specificity. These criteria accept deformity in lieu of current swelling. The United Kingdom family collection of affected sibling pairs with rheumatoid arthritis has therefore used the 1987 ACR criteria modified for epidemiological study (table 1.1.3) as the basis of disease diagnosis during recruitment.

1.1.4 Disease costs

RA is an important cause of chronic disability. Poor joint function and disability are initially caused by active disease where involved joints are too swollen and painful to undertake normal tasks. In the longer term, bony erosions leading to significant joint damage, tendon rupture and muscle wasting cause a permanent loss of joint function and subsequent disability. The cost of any disease can be described in many ways but areas common to all include morbidity, mortality and financial loss. By any criteria, RA is very costly disease.

1.1.4.1 Morbidity

Quality of life may deteriorate significantly following the diagnosis of RA and is affected by many factors. Swollen, painful, stiff joints lead to poor function and RA is an important cause of chronic disability. Some of these effects can be readily measured with a range of outcome measures developed for RA (see appendix 5.1 for some examples) but the consequences of poor function leading to loss of a working or a social life are less easily documented. Compared to disease-free individuals of the same age most have greater difficulty in carrying out household work, social relationships, leisure and recreational activities, work etc. (Felts and Yelin, 1989). In one longitudinal study functional disability was assessed in 1274 patients using the Stanford HAQ disability index (Wolfe *et al.*, 1988) (appendix) and Functional Disability Index (FDI) over a 12 year period. Within 2 years of diagnosis half the

Plate 1.1.3 shows the forefeet of an individual with severe RA. Bilateral hallux valgus can be seen. Both second toes over-ride the big toes. The MTP joints are subluxed and some of the toes show clawing as a result of fixed flexion at the PIP joints.

patients showed moderate loss of function, but this became severe within 6 years and very severe by 10 years. The progression of disability was most rapid in the early years and worsened more quickly in women (Wolfe and Cathey, 1991). A recent community-based study undertaken in Finland showed that RA was associated with more than a seven fold increased risk of disability compared with a general population of adults from the same community (Sokka *et al.*, 2003). The impact of disability due to RA was greater in younger and middle-age people than in elderly patients.

Factors indicating an unfavourable prognosis in RA were found to be uncontrolled polyarthritis, functional disability, erosions, positive rheumatoid factor, RA susceptibility or severity genes i.e. HLA-DR4 (DRB1* 0401, DRB1* 0404 etc.), the presence of extra-articular features and psychosocial problems.

1.1.4.2 Mortality

The Stanford HAQ disability index (Wolfe *et al.*, 1988) (appendix 5.1.6) is useful as a prognostic indicator of disease severity and length of survival. Another risk factor predicting greater disability and premature death is the requirement for an individual to be treated with prednisolone as this suggests greater disease severity. These prognostic factors predict a two-fold increase in mortality equivalent to malignant conditions such as Hodgkins Disease or chronic heart disease (Wolfe *et al.*, 1994). RA is also a marker for the development of co-morbid conditions such as renal disease. A recent population-based analysis of survival trends in RA over 40 years revealed that patients with RA were at significantly higher risk of death, with an standardised mortality ratio (SMR) of 1.27 (95% confidence interval 1.13 – 1.41). Excess mortality among women was more pronounced than among men (SMRs of 1.41 and 1.08 respectively) (Gabriel *et al.*, 2003).

Plate 1.1.6 is a radiograph of a hand showing the destructive changes of advanced RA. This individual has previously had a Darrachs procedure where the ulnar styloid has been removed surgically for pain relief and to avoid risk of subsequent extensor tendon rupture. The MCP joints are subluxed and the PIP joints and carple bones of the wrist show signs of significant destruction.

1.1.4.3 Finances and work

A systematic survey undertaken in the United States (US) in 1988 investigated the direct and indirect costs of major diagnostic categories such as musculoskeletal disease, cardiovascular disease and cancer. Musculoskeletal diseases cost up to 2.5% of US gross national product (GNP), equivalent to a severe recession (Yelin, 1998). In the UK, the estimated cost of secondary care and drug treatment for new cases of RA per year is roughly £21.8 million (1990-1991 prices) (Cooper *et al* 2000). The total direct cost of RA for 232,825 people in the UK with disease is estimated at £604.6 million (1992 prices) (McIntosh, 1996). Non-health service costs (incurred by the individual, their family and friends) account for a substantial proportion (86%) of the total costs associated with early inflammatory arthritis (Cooper *et al*, 2002). Costs are not equally distributed. Figures suggest that incurred costs for those in the 10^{th} percentile are < 2% per capital income, for those in the 90^{th} percentile costs are 40% higher than per capita income but for those in the 99^{th} percentile costs are four times as large as per capita income (Yelin, 1998). In a UK community-based study (NOAR) one third of RA patients had stopped work on the grounds of ill-health within two years of symptom onset (Barrett *et al*, 2000), 50% within 10 years and 90% had left work before retirement age (Yelin *et al.*, 1987).

RA is a common and costly disease with a two-fold increased mortality, significant morbidity and high financial costs. Learning more about factors predisposing to disease susceptibility and severity is, without doubt, important.

1.1.5 The synovium

The synovium is a continuous layer, which lines joints, tendon sheaths and bursae. It forms a non-adherent surface allowing movement between the various components of the joint and encloses a synovial space containing a small amount of hyaluronan rich fluid. It consists of two layers: the intima and subintima. Synovial fluid and matrix components are synthesised by specialised cells within the intima. Hyaluronan and adhesion molecules are synthesised by intimal fibroblasts and intimal macrophages carry receptors for immune complexes. Inflammatory disease targets the intima and the intimal cells interact with B-lymphocytes, antigen-antibody complexes and complement proteins. In health, macrophage numbers are comparatively low but can account for up to 80% of intimal cells in disease. The binding of immune complexes to intimal macrophages in RA releases cytokines and may induce synovial fibroblasts to support B cell survival and maturation (Edwards, 1998).

1.1.6 Inflammation

Inflammation is the response of living tissues to injury (Prockop and Kivirikko, 1995; Robinson, 1998). It is an ancient physiological response to insult and is an essential factor in the protection of the host from attack by pathogenic micro-organisms. Initially it is a destructive process, which should be limited by controlling anti-inflammatory processes, which lead to a healing phase. The response occurs in vascularised tissues enabling the delivery of essential molecules and cells to sites of inflammation in extravascular tissues (Robinson, 1998). In some instances inflammation will resolve with little tissue damage; in others it is accompanied by significant tissue destruction and repair is only partial (for example septic arthritis with articular cartilage

damage). Attempted repair follows repeated episodes of inflammation in conditions such as RA. Chronic inflammation involves many different mediators and cell types and is characterised by the presence of mononuclear cells such as macrophages, lymphocytes, plasma cells, and the proliferation of connective tissue fibroblasts.

The vascular response to injury results in increased temperature, erythema and swelling around joints that is the hallmark of RA (Prockop and Kivirikko, 1995). The process involves dilatation of arterioles and increased permeability of the microvasculature. Plasma then exudes into the extravascular space and leucocytes migrate through widened gap junctions and basement membrane into the extravascular space (Robinson, 1998). The loss of plasma increases blood viscosity. These events are induced and modified by a large number of inflammatory cells and the proteins produced by these cells. Protein synthesis and release depends upon the interaction of specific cell receptors and can result in both synergistic and antagonistic effects.

An inflammatory reaction requires the activation of different classes of leucocytes including, neutrophils, monocytes, macrophages, mast cells, eosinophils and basophils (Hardingham and Fosang, 1992). Some leucocytes are attracted to sites of inflammation, whereas monocytes reside in peripheral tissues and can differentiate into macrophages as a response to environmental stimuli. Macrophages are important in inflammatory reactions performing two important functions; phagocytosis and antigen presentation (Robinson, 1998). Following phagocytosis the ingested material is degraded to low molecular weight fragments and presented on the cell surface bound to class II MHC antigens (section 1.2.3). The antigenic fragments are recognised by T cells in relation to the MHC antigens. This leads to clonal expansion and activation of antigen specific T cells.

Lymphocytes can identify both foreign and auto-antigens and are involved in the responses designed to eliminate these antigens (Prockop and Kivirikko, 1995). B lymphocytes recognise antigenic epitopes and then proliferate and differentiate into antibody forming plasma cells. T cells recognise antigens through their T cell receptor and this also leads to proliferation. CD4 T cell subsets secrete lymphokines including interleukins IL2, IL4, IL5, which facilitate the amplification of B cell populations (Robinson, 1998). They augment delayed hypersensitivity reactions by enhancing cytotoxic T cells and facilitate the activation of macrophages by secreting interferon γ (IFNγ). They promote the growth and differentiation of bone marrow stem cells by secreting granulocyte-macrophage colony stimulation factor (GM-CSF). CD8 T cell subsets carry certain antigens in combination with class I MHC antigens (section 1.2.3) and are important in host defence against virally infected cells and the rejection of alloantigens. They are cytolytic to cells and generally suppress immune responses.

There are many mediators of inflammation including the complement system, kinins (such as histamine, adenosine, serotinin) nitric oxide, the clotting system and activated forms of oxygen as free radicals (Robinson, 1998). Prostaglandins (PG) and thromboxane A_2 including PGE_2 and PGI_2 are mediators of the vascular phase of inflammation being potent vasodilators. They also stimulate osteoclastic bone resorption, and bone erosions in RA may be partly mediated by them. Other mediators of inflammation, leukotrienes, platelet activating factors, and proteases, are all important mediators of tissue injury and degradation. In health these are tightly regulated to avoid uncontrolled tissue destruction and are generally stored within leucocyte lysosomes but some are synthesised de novo, and secreted in response to inflammation.

Cytokines are polypeptides secreted by a number of cell types, to regulate growth, differentiation and activation of leucocytes (Robinson, 1998). The

interleukins (IL) are a family of cytokines with varying functions, some being pro- and some anti-inflammatory. IL-1, tumour necrosis factor (TNF) and IL-6 have overlapping functions and are produced in large quantities by monocytes. They stimulate the synthesis of acute phase proteins by the liver, augment T cell growth and facilitate B cell proliferation and immunoglobulin secretion. IL-1 induces IL-6 secretion thereby promoting osteoclast activation. TNF acts synergistically with IFNγ to enhance antiproliferative effects. IL-2 is produced by T_H cells and stimulates the proliferation of T cells. IL-4 promotes T cell growth, can induce immunoglobulin class switching and increases the expression of class II MHC antigen on macrophages and B cells. IL-3 promotes the growth and differentiation of haematopoietic stem cells. IL-5 facilitates the growth and function of B cells and eosinophils. IL-7 is a growth factor for T cells and IL-8 causes chemotaxis and activation of neutrophils. IL-10 is considered to have mainly immunosuppressive functions but increases the differentiation and activation of B cells (section 3.1.2).

Inhibitors of cytokine function play a major role in normal immunity and homeostasis by interrupting positive feedback loops and restraining cellular activation. There are four basic mechanisms of cytokine inhibition including soluble receptors (e.g. soluble IL-2 receptor) receptor-binding proteins such as IL-1 receptor antagonist (IL-1ra) (Arend and Guthridge, 2000), anti-cytokine antibodies (e.g. anti -IL-1α) and immuno-suppressive cytokines such as IL-10 and IFN-α.

1.1.7 Pathology of RA

Chronic synovitis is one of the central features of RA although the pattern of synovial involvement varies from case to case. Initially there is evidence of hypertrophy of the lining layer of synovium with infiltration by activated macrophages and lymphocytes and proliferation of blood vessels. $CD4^+$ (T

helper) lymphocytes predominate in the peri-vascular areas and are in close proximity to antigen presenting cells. B lymphocytes also infiltrate and produce rheumatoid factors (RF) (Firestein, 1998). These are auto-antibodies directed against the Fc fragment of immunoglobulin. The synovium develops as an invasive pannus which erodes the articular cartilage and bone especially where the synovium attaches to the joint margin (Firestein, 1998). This cartilage and bone destruction leads to the loss of joint space and erosions evident on X-ray (plates 1.1.5 and 1.1.6). The joint becomes painful and an effusion leads to joint swelling (plate 1.1.1). Large numbers of cells such as neutrophils, macrophages and lymphocytes are found in the synovium and synovial fluid. These cells produce pro-inflammatory cytokines such as IL-1, TNFα and IL-6, chemokines such as IL-8 and growth factors including transforming growth factor β (TGF-β), platelet derived growth factor (PDGF) and fibroblast growth factor (FGF). The acute phase response, stimulated by IL-1, IL-6, TNFα, is pronounced with high levels of circulating C-reactive protein, serum amyloid protein and polyclonal hypergammaglobulinanaemia.

Radiographs of affected joints show soft tissue swelling, juxta-articular osteoporosis and joint space narrowing (Bower, 1998) (plates 1.1.5 and 1.1.6). Cartilage is eroded and this is followed by the development of bone erosions and joint deformity leading to instability. Bone re-absorption is seen early, partly reflecting disuse but also local cytokine release. The metalloproteinases are responsible for matrix degradation and remodelling. They are generally secreted as inactive proenzymes by synovial lining cells, requiring proteolytic cleavage for activation. Their secretion is mediated by other cytokines. IL-1 and TNF-α are potent inducers whilst interferon γ (IFN-γ) decreases activity. Tissue inhibitor of metalloproteinase (TIMP) is a key inhibitor and blocks activity by binding to the metalloproteinase. Stromelysin, and collagenase are two of the most important metalloproteinases because in combination they can degrade virtually all important structural proteins within

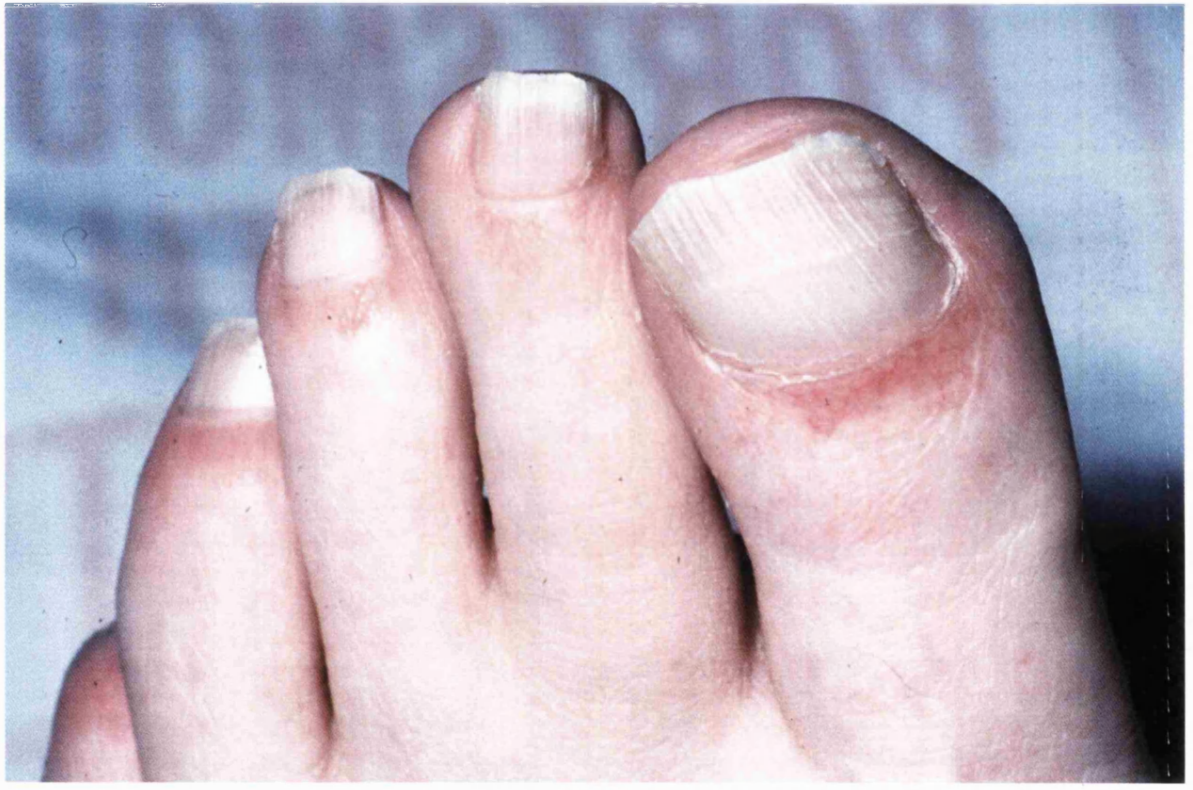


Plate 1.1.9 shows the toes of an individual who has RA and a nailfold vasculitis. This suggests that RA vasculitis may be developing and additional treatment may be required.

a joint. Stromelysin digests interstitial collagenase, cleaves proteoglycans, laminin and fibronectin and activates latent collagenase which in turn digests native triple helical collagen leading to matrix destruction (Firestein, 1998). Stromelysin mRNA and protein has been identified in synovial tissues early in the disease process and collagenase levels are higher in RA synovial tissue than osteoarthritis (OA) (Firestein *et al.*, 1991). A long term study of low dose methotrexate (section 1.1.9), a drug shown to reduce the progression of bone erosions in RA, showed a 70% decrease of collagenase mRNA but no change in stromelysin or TIMP expression (Firestein *et al.*, 1994)

Prostaglandins are synthesised by cells within the synovial fluid and synovium. Prostaglandins (for instance PGE_2) potentate both vasodilation and the changes in vascular permeability induced by complement fragments, histamine and leukotrienes. PGs can also inhibit the production of anti-inflammatory cytokines such as IFN-γ. The immune complexes within a joint affected by RA can lead to complement activation. Rheumatoid factors (RF) are a consistent component of the immune complexes and serum rheumatoid factor levels correlate positively with complement. Complement components are mainly synthesised in the liver but can also be produced locally within the joint.

Human leucoctye antigens (HLA) associations have been reported in a number of rheumatic diseases but only certain HLA-DRB1 alleles have been associated with RA (section 1.2.3). Peptide is presented to T cells by antigen presenting cells in conjunction with class II MHC antigens (section 1.1.6). There are a number of plausible explanations (Weyand and Goronzy, 1995) regarding the mechanism of the class II association and these are discussed in section 1.2.3.

Prediction of RA severity: clinical patterns of the disease course

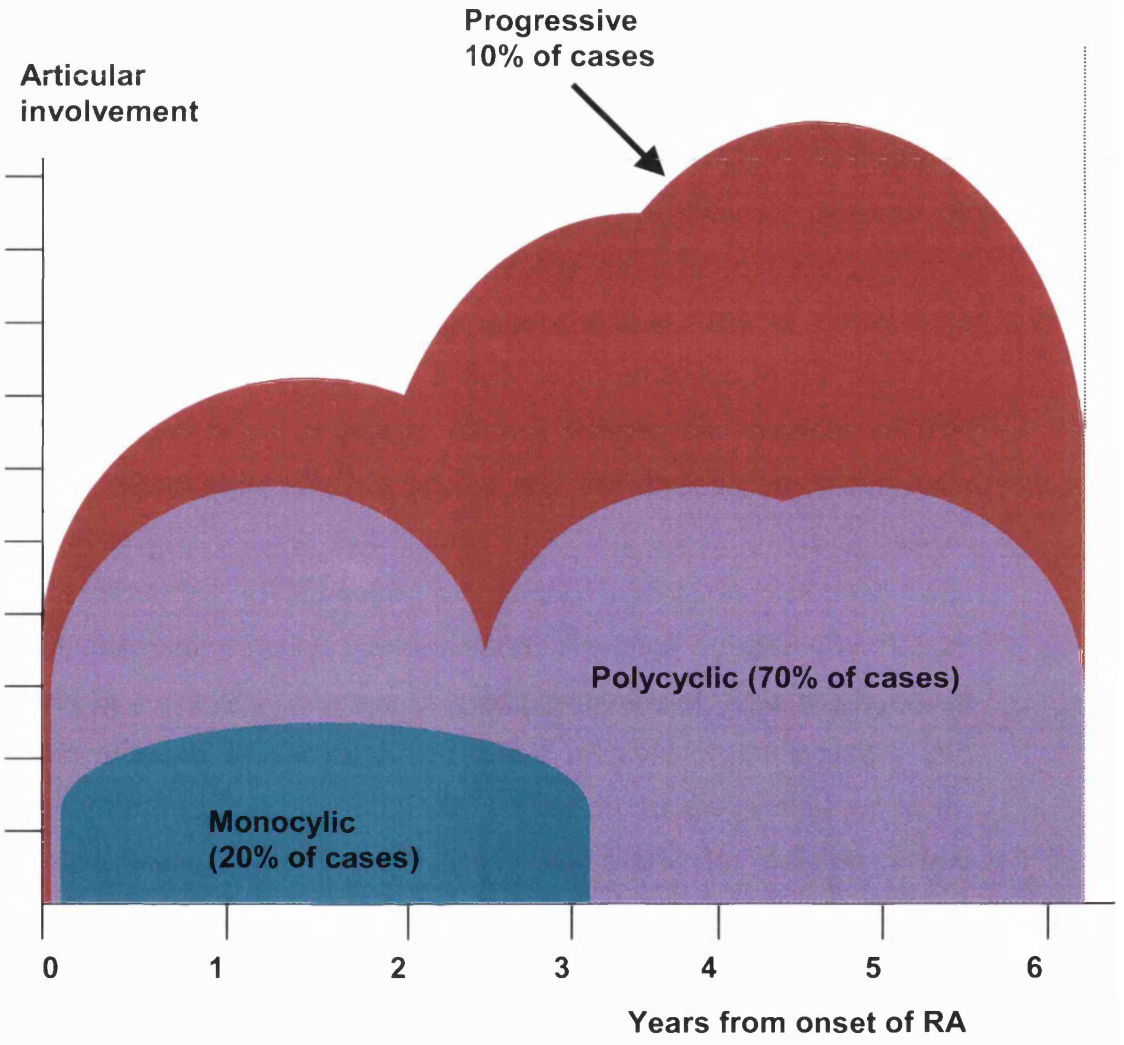


Fig 1.2 adapted from Gordon and Hastings [Gordon, 1998 illustrates the various patterns of the clinical disease course in RA. The majority (70%) of individuals with RA follow a polycyclic pattern of disease activity where disease activity is either intermittent with short periods of remission or continuous with differing levels of severity. Up to 20% of individuals have a monocyclic pattern of disease activity with a single cycle followed by remission for at least one year. Approximately 10% of people follow a progressive disease pattern with increasing joint involvement and no sign of remission.

1.1.8 Predicting disease severity

Predicting disease severity in any individual is difficult. Prognosis may be associated with differences in presentation. Those who experienced an acute onset of disease may have a better functional outcome than those presenting with insidious progressive disease but this is by no means clear cut (Fleming *et al.,* 1976; Fleming *et al.*, 1976; Jacoby *et al.*, 1973). Additionally, subsequent disease progression may follow several different patterns off disease presentation. Three articular patterns have been described and include a monocyclic pattern affecting approximately 20% of cases who experience a single cycle of disease activity with remission for at least one year. The second is a polycyclic pattern where the majority (~70%) experience an intermittent disease course and the disease smoulders with incomplete remission in some individuals. The third is a progressive pattern where approximately 10% have progressive joint involvement with uncontrolled disease activity (Gordon and Hastings, 1998) (fig 1.1.1). However, RA is a systemic disease despite predominately affecting synovial joints. Extra-articular involvement can affect multiple organ systems and systemic features may be associated with a poor prognosis (Vollertsen *et al.*, 1986). To accurately predict who will develop extra-articular features or who will continue with smouldering or uncontrolled disease activity is at present impossible.

The development of better ways to predict who will have a poor outcome early in the disease course, before significant joint damage has occurred, is of great importance. If this could be achieved, patients with a poor prognostic outlook could be treated aggressively with drug regimes that would be inappropriate to give to others.

Valid measures of outcome are necessary to follow the disease natural history and to measure success or otherwise of treatment. Two principal outcome measures are the development of radiological erosions and functional disability. Both have some correlation to long term outcome (see section 1.1.4.1). A number of studies have investigated the predictive powers of screening for HLA-DR4 or the 'shared epitope' (section 1.2.3) in those with early RA (Emery and Salmon, 1995; Reveille *et al.*, 1996; Weyand *et al.*, 1992). Although some evidence supports the association of HLA-DRB1 alleles with radiological erosions (Gough *et al.*, 1994; Harrison *et al.*, 1999; MacGregor *et al.*, 1995; Reveille *et al.*, 1996), the Norfolk Arthritis Register (NOAR) study has shown that the 'shared' or 'rheumatoid' epitope has only a modest effect on functional disability (Harrison *et al.*, 1999). Other studies have found no improvement with predictions of severity associated with knowledge of HLA-DRB1 status (Drossaers-Bakker *et al.*, 2002; Eberhardt *et al.*, 1996). Combinations of clinical and laboratory measurements taken early in the course of the disease have been studied to show whether it is possible to predict future disease severity. Levels of rheumatoid factor, number of bony erosions, patient's age and the number of swollen / tender joints at diagnosis (+/- the HLA-DRB1 antigen) have been shown to partially predict the outcome of the disease process (Drossaers-Bakker *et al.*, 2002).

1.1.9 Treatment

The current strategy for the treatment of RA is to fully suppress disease activity as quickly as possible (Wilske, 1993). A number of disease modifying drugs (DMARD) are available to achieve this end. These include methotrexate, sulphasalazine and intra-muscular gold (Jeurissen *et al.*, 1991; Rich *et al.*, 1999; van der Heijde *et al.*, 1989). These drugs are used singly, or in combination if disease suppression is not achieved using a single DMARD (O'Dell *et al.*, 1996; Tugwell *et al.*, 1995; Wilske, 1993). These drugs are potentially toxic and require careful monitoring. They also take some

months to have an effect upon the disease process. However, despite the time taken to settle the joint symptoms and to reduce the acute phase response evidence is available to show that radiographic progression of joint disease can be slowed (Pincus *et al.*, 2002). Function should be preserved if joint destruction is limited. Non steroidal anti-inflammatory drugs (NSAIDs) are often used in conjunction with DMARDs to help treat day to day symptoms of pain and stiffness. Alternatively low dose corticosteroids are used in conjunction with DMARDs. There is evidence that these have some disease modifying effect and reduce the occurrence of erosions (Kirwan, 1995) but there are many potential serious side effects and these drugs need to be used with caution. In recent years biologic therapies have become available, in particular anti-TNF therapies (Maini *et al.*, 1999; Weinblatt *et al.*, 2003; Weinblatt *et al.*, 1999). This approach has improved disease control in many patients who had failed conventional DMARD treatment. They have slowed radiological progression (Arend, 2002; den Broeder *et al.*, 2002). Additionally biological therapies appear to have a more profound effect on many of the constitutional symptoms such as fatigue than conventional therapy (Arend, 2002). Short-term side effects appear to be fewer and less rigorous monitoring of liver and marrow function is required than for some conventional DMARDs. However these drugs have only recently become available and the longer-term side effects remain unknown. Particular concerns remain with respect to infections such as tuberculosis and the risk of malignancy.

1.1.10 Iatrogenic effects of treatment

The treatments used to control rheumatic disease frequently cause adverse drug reactions or unintended side effects. These can be divided into two groups: the relatively predictable 'toxic' reactions and the unpredictable idiosyncratic reactions. The former are usually dose-dependent and related

to the intrinsic reaction of the drug whereas the latter are usually immune-related. Prevention of these adverse events is very important. Before treatment with DMARDs is commenced patients should be screened for contra-indications such as pre-existing liver disease, potential drug interactions or pregnancy. Adverse drug reactions are more likely to occur early in the course of treatment therefore monitoring should be carried out more frequently during the first few months of treatment. Table 1.1.7 lists the common adverse effects of the various DMARD treatments currently available for RA. A substantial minority of patients develop side effects to many ranging from mild effects such as nausea to more serious problems including renal or liver disease, severe rashes or pulmonary fibrosis. There are concerns regarding the long-term use of some DMARDs and the potential to develop a malignancy, particularly a B cell malignancy (Wijnands and van Riel, 1995). This is a particular concern when considering the biological therapies as long-term data is not yet available regarding the risk of developing a malignancy or serious infection.

The adverse effects of long-term corticosteroid treatments are well recognised and depend upon both the size of the dose and the length of treatment. Monitoring (± treatment) for hypertension, diabetes, osteoporosis, peptic ulceration, cataracts, and steroid induced myopathy is important because all are recognised complications of steroid treatment. However, if steroid doses are kept to a minimum and strategies of pro-active treatment are instituted in those at particular risk of a complication (e.g. bone protection treatments in those at a higher risk of osteoporosis) these drugs can be very effective at reducing disease activity. Steroids can be given orally, by intra-muscular injection, by intra-articular injection or intravenously in the form of a high dose 'pulse' of treatment used to gain prompt control of disease activity. Each route can lead to particular adverse reactions. Daily doses of oral corticosteroids or frequent intra-muscular administration may cause those side effects mentioned above. Intra-articular steroids are frequently used to

control local synovitis, producing prolonged relief of joint symptoms and improved function. Potential adverse events include the introduction of intra-articular sepsis and concerns that too frequent administration may lead to bone and cartilage damage (although little data is available to support this view). Intravenous administration can lead to facial flushing, psychological disturbances, palpitations and occasionally severe adverse reactions such as cardiac arrest or severe infections in those with failing cardiovascular or immune systems. However, intravenous administration of corticosteroid is usually free of longer-term side effects.

NSAIDs are very commonly prescribed for those patients with any form of arthritis or musculoskeletal pain. They are almost always prescribed to individuals with RA at some stage in the disease course. For a few with very mild disease or arthritis in remission a NSAID is sufficient to control disease activity but usually they are used as an adjunct to DMARD therapy. The commonest adverse events are related to the gastrointestinal (GI) system and include dyspepsia, peptic ulceration or perforation and gastrointestinal bleeding. Reports suggest that between 2-300 individuals per year die in the UK from NSAID induced GI side effects (usually peptic ulceration, perforation or bleeding) (Blower *et al.*, 1997). The anaemia an individual with chronic inflammation is likely to develop is normochromic and normocytic, hence a microcytic, hypochromic anaemia usually indicates GI bleeding (possibly related to NSAID treatment). Other adverse events associated with NSAID therapy include bronchospasm in asthmatics, abnormal liver function tests, rashes, renal impairment and diahorroea.

The range of adverse events, both predictable and idiosynchractic, associated with treatment for RA emphasises how important it is to understand the natural history of disease more fully and to be able to identify those who may be at a greater risk of severe disease. If this were possible it would be easier to avoid treating those with mild disease with potentially toxic

Drug	Adverse effects - not influenced by dose	Adverse effects - influenced by dose	Adverse effects - not influenced by cumulative dose	Adverse effects -influenced by cumulative dose
Methotrexate	Pneumonitis Hepatotoxicity	GI side effects Stomatitis Raised transaminases	Pneumonitis	Liver fibrosis
Sulphasalazine	Leucopaenia Rashes Hepatitis	GI side effects Headaches Haemolysis		
Parentral gold	Rashes Haematological	Rashes Proteinurea Diarrhoea		Chrysiasis
Azathiorpine	Nausea	Leucopaenia Hepatotoxicity		
Leflunamide	Diarrhoea	Diarrhoea Hypertension		
Hydroxy-chloroquine	Keratopathy	Retinopathy GI side effects Muco-cutaneaous	Rashes GI side effects Muco-cutaneaous	Retinopathy
Oral gold	Rashes GI side effects	Diarrhoea		
Cyclophos-phamide	Diarrhoea Nausea	Alopecia Leucopaenia		Heamorrahagic cystitis Malignancy Gonadal dysfunction
D-penicillamine		Haematological Renal		
Corticosteroids		Osteoporosis Glucose intolerance Infection		Osteoporosis Myopathy Cataracts
Etanercept	? Infection injection site reactions			? malignancy
Adalumimab	? Infection injection site reactions			? malignancy
Infliximab	? Infection			? malignancy
Anakinra	? Infection			? malignancy

Table 1.1.7 adapted from Wijnands and van Riel (Wijands and van Riel, 1995) illustrates the relationship between the dose of a DMARD and the adverse events that may occur when a patient is treated with that particular DMARD. Some adverse events are more likely to occur depending upon either the current or the cumulative dose of the DMARD whereas others are idiosyncratic and may occur at any time. As the biological therapies are relatively new forms of treatment the effect of a cumulative dose can only be guessed at.

drugs. It may also be possible to predict individuals who are likely to develop side effects when treated with certain drugs. Some progress has been made in this area and certain tests can be performed prior to starting some DMARD therapies such as testing for thiomethylpurine transferase deficiency prior to starting azathioprine (Corominas *et al.*, 2003; Marra *et al.*, 2002). Another test that can be undertaken prior to starting methotrexate is the level of the N terminal fragment of procollagen III (P3NP) (Mitchel *et al.*, 1990). These tests can help predict individuals who are more likely to develop some of the more serious adverse effects associated with these particular therapies. However, greater progress is needed in these areas.

1.1.11 Conclusion

Rheumatoid arthritis is a common disease associated with significant morbidity, disability, financial cost and an excess mortality. It is clinically heterogeneous and although diagnostic criteria and outcome measures are available it can be difficult to adequately categorise patients in terms of disease severity or prognosis. This can lead to difficulties in determining the extent (or aggressiveness) of initial treatment and which individuals to include in drug or genetic studies. Ultimately it may be necessary to abandon the concept of RA as a single disease entity. Distinct RA subsets could then be identified and patients stratified into categories that differ with respect to aetiology, disease course, clinical pattern and treatment response. Understanding more about the causes of RA, disease pathogenesis and predictors of severity is very important. Onset of disease is likely to occur when a genetically susceptible individual encounters some form of trigger (which may be quite a ubiquitous trigger). A greater understanding of the genes contributing to genetic susceptibility or even disease severity should improve our understanding of the natural history of RA, may allow the identification of those who require aggressive treatment from an early stage and may help in the development of novel therapies in the future.

Chapter 1 – section 2

CHAPTER 1 - section 2

Genetic principles and techniques

1.2.1 Introduction

Rheumatoid arthritis has a complex or multi-factorial inheritance in common with many diseases and the majority of non-pathological traits such as height, skin and eye colour. Disease onset is thought to occur when an individual with a susceptible genetic background encounters an environmental trigger or triggers (Nepom and Nepom, 1998). Both need to be present for the disease to develop and progress. However, it is important to recognise that non-genetic factors do not simply consist of exposure to an environmental influence such as infection but also include processes that may occur early in development (Seldin *et al.*, 1999). The environmental trigger(s) may be unknown but some evidence does exist as to the genes involved in susceptibility. As a result of this rather complex interplay between genes and environment, any genetic associations with RA are unlikely to be a simple one to one correlation between a gene and a phenotype. The relationship is likely to be obscured by a considerable level of noise and will only be resolved by undertaking a large number of observations (Seldin *et al.*, 1999).

1.2.2 Genetic predisposition to RA

Evidence regarding the genetic predisposition to rheumatoid arthritis is available from a number of sources. These include epidemiological studies documenting prevalence rates amongst the family members of a proband (familial aggregation), twin studies and case control studies showing association between rheumatoid arthritis and certain HLA-DRB1 alleles (shared epitope hypothesis, section 1.2.3). Variation in disease severity is

also due (in part at least) to genetic factors. For instance, the number of erosions documented on a hand x-ray (one predictor of severity) is associated with the MHC class II genotype (Eberhardt *et al.*, 1996; Harrison *et al.*, 1999; MacGregor *et al.*, 1995; Reveille *et al.*, 1996).

1.2.2.1 Twin and family studies

Family studies show that the first-degree relatives of probands with RA are at increased risk of developing RA and that this is particularly true if the proband has severe RA (Lawrence, 1970). This lends support to the hypothesis that the genetic background of an individual contributes to RA pathogenesis. The relative recurrence risk ratio (λ) is defined as the prevalence of the disease in the target group such as the siblings of probands (K_s) divided by the prevalence of the disease in the general population (K). Thus $\lambda_s = K_s / K$. First degree relatives (with 50% gene sharing) such as offspring, parents or siblings are at greatest risk; this risk is reduced for second degree relatives (e.g. aunt, niece, grandchild); and less closely related individuals (e.g. first cousins) have a risk very near that of the general population (fig 1.2.1). This implies that a limited number of genes are required to have a genetic predisposition to RA and second-degree relatives are much less likely to have inherited the full complement of genes required in comparison to first-degree relatives. A recurrence risk in a complex disease is proportional to the population risk of the phenotype. This has relevance when studying a fairly common disease such as RA (section 1.2.2.2). Determining disease prevalence is subject to various methodological difficulties (section 1.1.2.1) and it is preferable to estimate disease prevalence and familial aggregation in the same population, using the same criteria (section 1.1.3). The disease severity in the proband will also affect the sibling recurrence risk (λ_s). Where the proband has mild non-erosive sero-negative disease λ_s is barely greater than unity but it increases in the siblings of patients with sero-positive erosive

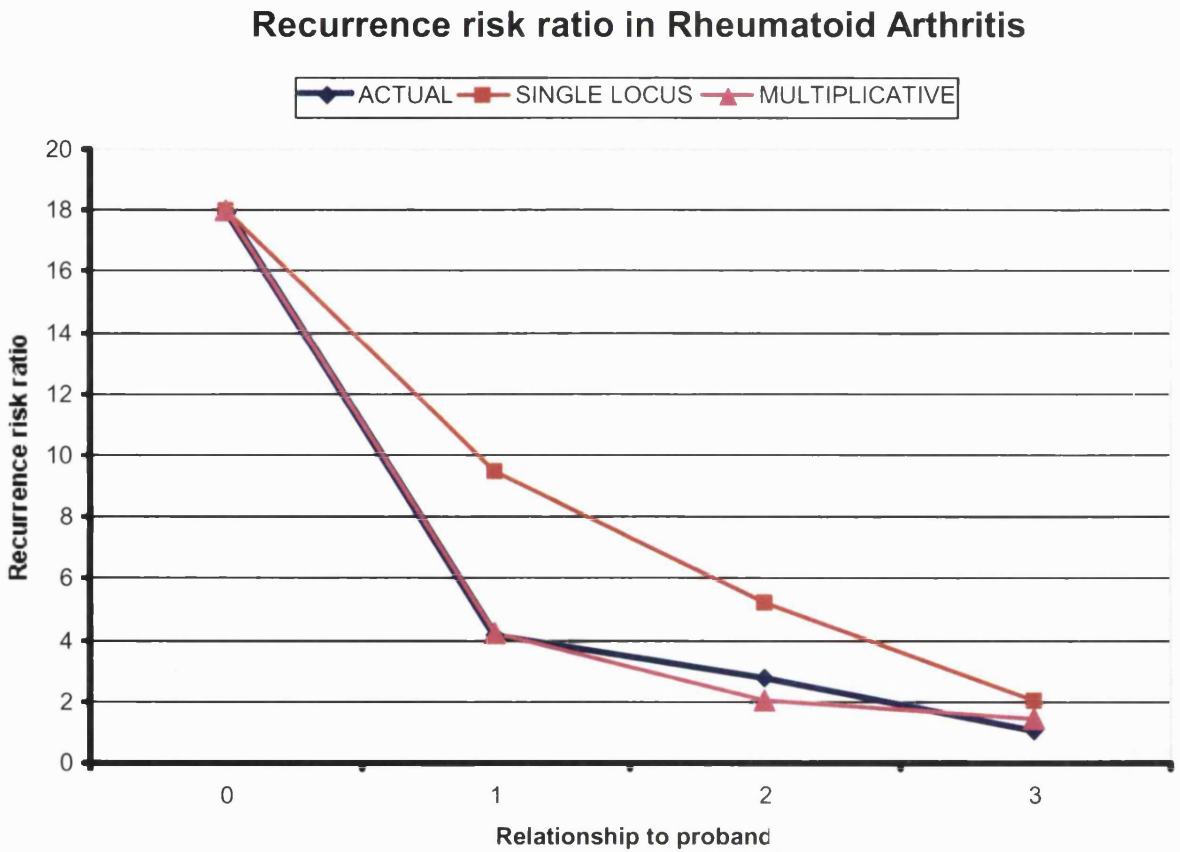


Fig 1.2.1 illustrates a recurrence risk curve drawn for RA using UK prevalence data. It can be seen that the recurrence risk falls sharply with increasing distance of relationship to the proband. Although much of the risk reduction seen with less closely related individuals is thought to be due to a difference in the extent of genetic similarity it is also likely that first-degree relatives will share more common environments than those who are more distantly related. The red line indicates the likely curve for a single locus model and the pink line the likely curve for a multiplicative model.

disease (Lawrence, 1970). The recurrence risk ratio in the siblings of probands with severe disease may be as high as λ_s = 15 suggesting that genes are involved in determining disease severity (Wordsworth and Brown, 1998). This is because the prevalence of severe RA in the general population (K) is lower than with milder forms of RA while the recurrence risk ratio is higher. Siblings of RA probands who share both HLA haplotypes with the proband are at particular risk of developing the disease (Nepom and Nepom, 1998).

Concordance rates in monozygotic twins vary between 12 and 30% but different methodologies have been employed in the studies. Lawrence studied twins with sero-positive erosive disease and obtained a disease concordance figure of 30% (Lawrence, 1970). This contrasts with the lower rate of 12% reported by Aho in a nationwide survey of twins, which included individuals with less severe disease (Aho *et al.*, 1986). A United Kingdom survey found an overall figure of 15% but reported that concordance rates were highest in monozygotic twins who were HLA DR4 positive (Silman *et al.*, 1993). These figures represent the results from cross-sectional studies in a disease where it is known that there may be a delay of up to 40 years between the development of the disease in the proband and that of the second twin. Hence there are problems in assessing the heritability.

1.2.2.2 Heritability

The study of twins has been the classical method of determining the extent of genetic and environmental influences on quantitative traits and disease predisposition. However, interpreting the results of these studies is often not straightforward. Rheumatoid arthritis is a good example. The MZ twin disease concordance from two large twin studies was between 12 - 15% (Aho *et al*, 1986; Silman *et al*, 1993). This is often incorrectly assumed to indicate

that the predominant influence in disease predisposition is environmental and the genetic contribution to disease is low. However, absolute levels of concordance are dependent on the population prevalence of a disease [MacGregor *et al*, 2000]. As the prevalence of a disease increases so too does the twin concordance, irrespective of the genetic contribution. Therefore twin concordance figures provide only a limited insight into the overall contribution of genetic factors. The estimation of 'heritability' is more readily interpretable and can better gauge the relative genetic and environmental contributions. 'Heritability' estimates the extent to which the variation in the population's liability to disease can be explained by genetic variation. It is independent of disease prevalence and a quantitative estimate of the genetic influence. The estimate of RA heritability is in the region of 60% (MacGregor *et al*, 2000).

Twin studies are unique in that they provide an estimate, unconfounded by age, of the extent of genetic influences on disease predisposition which can be estimated separately from the influence of the shared family environment. The magnitude of heritability determines the power and hence the likely success of programmes designed to identify specific genes involved in the disease (MacGregor *et al*, 2000). It also provides a yardstick against which to assess the relative importance of candidate genes that may be identified as being associated with disease. The genetic contribution to rheumatoid arthritis may not be constant across all disease subgroups. Rheumatoid arthritis is more prevalent in females, and its incidence and severity may be declining over time (Jacobsson *et al*, 1994; Linos *et al*, 1980). RA is more prevalent in families with other autoimmune diseases such as autoimmune thyroiditis or Type 1 diabetes mellitus suggesting the presence of common autoimmune susceptibility loci (Myerscough *et al*, 2000). The strength of the genetic association with HLA is reported to vary according to gender and disease severity phenotype and may also vary in different birth cohorts [Silman *et al*, 1983]. An understanding of the extent of the differences in

The human leucocyte antigen (HLA) complex

Fig 1.2.2a

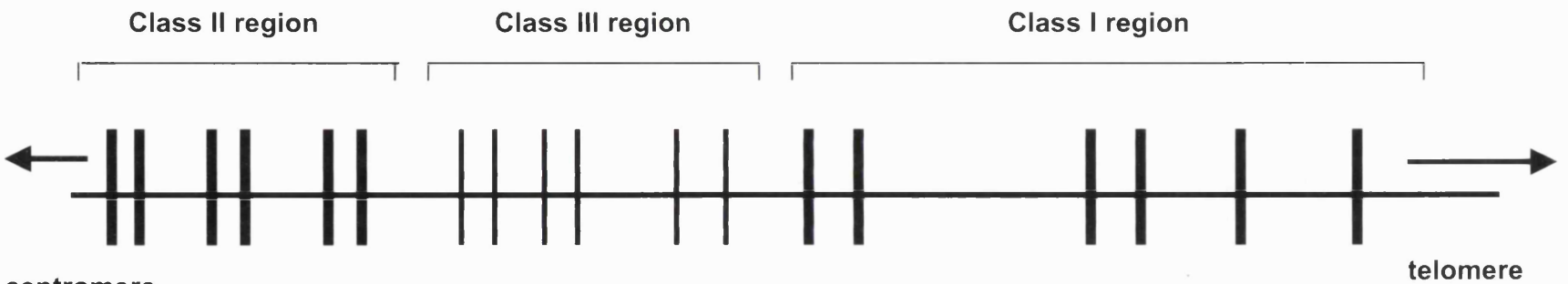
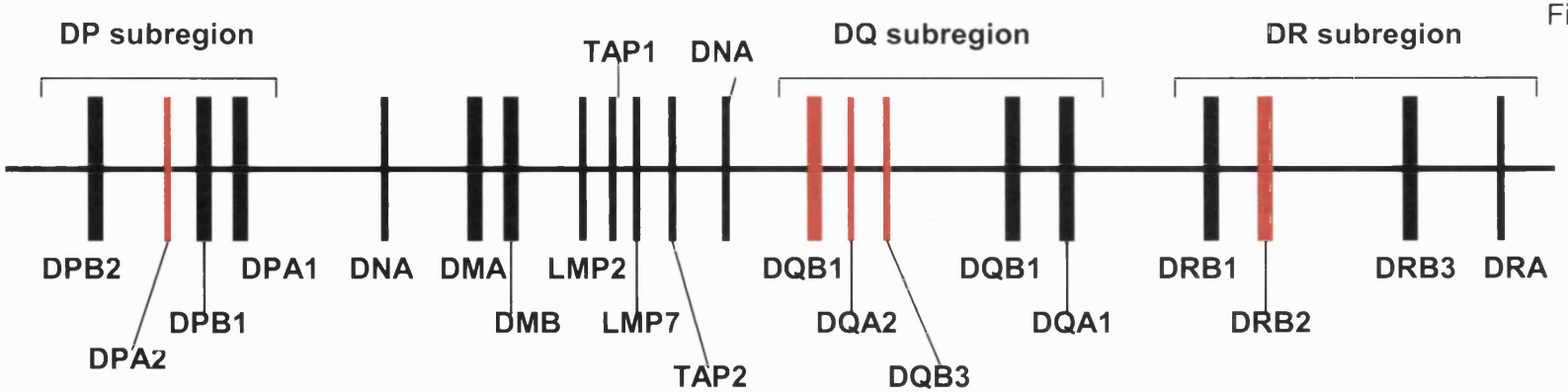


Fig 1.2.2b



Functional gene

Pseudogene

Fig 1.2.2a illustrates the class 1, 2 and 3 regions of the human leucocyte antigen complex (HLA) on chromosome 6. An expansion of the class 2 region is shown in **fig 1.2.2b** where functional genes are shown in black and pseudogenes in red. Genes within the DR, DQ and DP regions are shown. Subtypes of DRB1 have been associated with RA. This is an example of an HLA-DR3 haplotype which has DRB1 and DRB3 loci.

genetic contribution between sexes and between those with different disease characteristics is therefore important if subgroups with increased genetic risk are to be targeted for linkage and association studies.

1.2.3 The Major Histocompatibility Region (MHC)

The major histocompatibility region (MHC) region maps to the short arm of chromosome 6 (6p21.3) and contains a cluster of genes governing immune reactions called the human leukocyte antigen (HLA) complex (Figs 1.2.2a and b). HLA class I genes, located at the telomeric end of the MHC region, include HLA A, B and C loci encoding the main transplantation antigens in humans. These are expressed on all nucleated cells and are highly polymorphic due to a high degree of allelic variation at each locus. The HLA class I heavy chain complexes with β2 microglobulin to form the mature HLA class I molecule. HLA class II genes are expressed on specialised antigen presenting cells including B cells, activated T cells, dendritic cells and macrophages. They are also highly polymorphic clustering into three main subregions, HLA-DR, DQ and DP. The mature class II HLA-DR molecule is formed from α and β polypeptides encoded by the HLA-DRA and DRB1 loci. The HLA-DQ and DP molecules are formed from the polypeptides encoded by HLA-DQA1 and DQB1 and DPA1 and DPB1, respectively

HLA class I and class II genes represent fewer than half of the known genes lying within the 4 megabases (Mb) of the human MHC. The HLA class III region lies between class I and II encoding, among other genes, the complement components C2, C4 and Bf. The TAP and LMP genes lie between the DQ and DP clusters encoding polypeptides implicated in HLA class I antigen processing and peptide processing. Other genes in this region with immunological functions include heat shock protein 70 (hsp 70) and tumour necrosis factor (TNF) (fig 1.2.2b).

Most HLA alleles at a single locus have highly conserved nucleotide sequences with one or two specific sites of diversity, mainly in the 5' proximal exons. These areas encode segments of the HLA molecule situated furthest from the cell membrane and include the recognition domains of the HLA molecule consisting of a deep groove or pocket. The floor of the pocket is made of β-pleated sheets and the sides α-helical loops. Most of the amino acid variation that distinguishes different HLA alleles occurs within this region because any amino acid variation can alter the properties of the antigen binding groove.

The main function of HLA molecules is to bind antigenic peptides and present them to T-cells, a process which allows the discrimination between self and non-self. Binding of the peptide depends upon the characteristics of the antigen binding groove which incorporates a number of pockets that accommodate the side chains of so-called anchor residues in the peptide (Doherty and Nepom, 1998). Differences in the amino acid composition of these pockets may have important effects on the array of peptides bound by a specific HLA molecule. An adequate 'fit' allows permissive binding to take place leading to antigen presentation and an immune response (if T cells are exposed to a 'foreign' antigen). A poor structural 'fit' means no binding, no presentation and no immune response. Hence, only those peptides able to bind to an HLA molecule have the potential to activate the immune response. With individual variation between HLA molecules, a peptide antigenic in one individual may not elicit a response in another if binding between the HLA molecule and peptide does not take place.

Antigen presentation by the class I pathway is distinct from that of the class II pathway. Cytosolic proteins (self and non-self) are proteolytically cleaved into peptides and bind to HLA class I molecules in the endoplasmic reticulum (ER). The class I-peptide complex is then transported to the plasma

membrane where both the peptide and class I molecule are essential for binding and activation of an antigen-specific T cell. There is increased affinity between T-cell and antigen presenting cell when the T-cell co-receptor, CD8, binds to the class I molecule. This facilitates activation and accounts for the apparent 'restriction' of $CD8^+$ cytotoxic T cells for MHC class I-associated antigens.

The Class II molecule is stabilised during formation in the ER by the invariant chain which blocks the peptide binding groove. This prevents association of peptide with the class II molecule whilst the molecule is in the ER. The class II-invariant chain complex then forms an acidic endosome outside the Golgi apparatus and partially digested 'foreign' material is deposited within this endosome. The acidic environment activates proteases to cleave the invariant chain allowing the 'foreign' antigen to bind to the class II molecule in the antigen binding groove. The class II-peptide complex then moves to the cell surface where it can be recognised by the T-cell receptor (TCR) (fig 1.2.3). Affinity between the TCR and class II-peptide complex is increased by the antigen CD4 binding with the class II molecule. This accounts for the class II / $CD4^+$ cell restriction.

Linkage disequilibrium (section 1.2.10.1) is common in the HLA region. Hence inheritance of a specific allele at one locus is frequently accompanied by the inheritance of a specific allele at another. Therefore if a particular HLA marker is associated with a certain disease the actual susceptibility gene may be the HLA marker itself or another nearby gene in linkage disequilibrium with that marker (i.e. different alleles tend to exist as haplotypes and these haplotypes are relatively fixed within a population). HLA associations have been reported with a number of rheumatic diseases but it is possible that

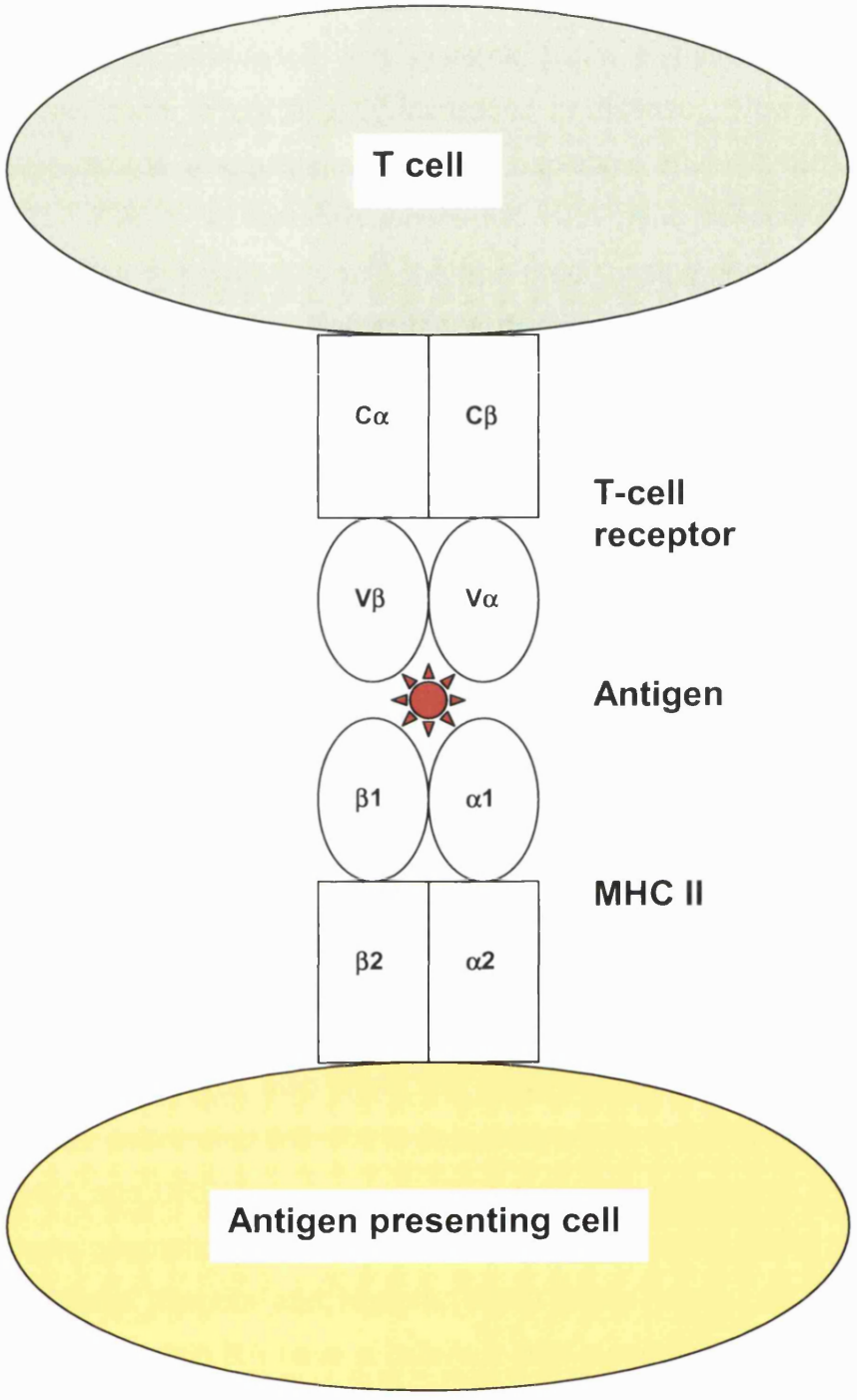


Figure 1.2.3 illustrates the association of the T cell receptor molecule, antigen and MHC class II molecule. Antigens are recognised by the α and β chains of the MHC which are expressed on the surface of the antigen presenting cell and the V regions of the T cell receptor.

different genetic mechanisms are involved with these associations. For instance, HLA-DR3 is associated with systemic lupus erythematosis (SLE) and the C4A null allele is significantly increased in disease. However, the C4A gene is in linkage disequilibrium with the haplotype HLA-A1, B8, Cw7, C4A and DR3. This is an example where the HLA gene associated with disease is in linkage disequilibrium with the true predisposing gene. In some cases, two HLA genes interact. For instance the risk of developing RA with the genotype HLA-DRB1*0401 and DRB1*0404 is 1 in 7 whereas the risk of disease is only 1 in 35 for the genotype DRB1*DRB1*0401 / X (where DRX is not DRB1*04 or DRB1*01) and 1 in 20 for the genotype DRB1*0404 / X (Nepom and Nepom, 1998).

A number of different HLA-DRB1 molecules have been associated with RA. However, only certain HLA-DRB1 alleles are associated and these differ in different geographical locations despite the few specific structural or genetic differences between the DR4 alleles. The shared epitope hypothesis (SE) explains these variable results (Gregersen *et al.*, 1987). HLA-DRB1*0401 and DRB1*0404 are the most commonly associated with RA (Wordsworth *et al.*, 1989) and these alleles carry the sequence LLEQKRAA or LLEQRRAA between codons 67 to 74 of the DRβ molecule. HLA-DRB1*0101, DRB1*0405 and DRB1*1402 also share the same sequence as DRB1*0404 at this site and are associated with RA in populations with a low prevalence of HLA-DRB1*0401 and DRB1*0404 genes. DRB1*0405 is associated with RA in the Japanese population, DRB1*0101 with Israeli Jews and DRB1*1402 with Yakima Indians (Nepom and Nepom, 1998) (table 1.2.1). HLA-DRB1 genes not associated with RA have a different DNA sequence in this region frequently including negatively charged amino acids such as aspartic acid at position 70 or glutamic acid at position 71 nor 74. This area between codons 67 to 74 forms part of the antigen binding groove and the fourth binding pocket; the size and shape of the groove could be altered by a positively or negatively charged amino acid hence determining peptide binding (Nepom

HLA-DR associations with RA amongst different ethnic groups

HLA-DR type	Subtypes	Strength of association	Nationality
HLA-DR4	*0401	++	N. European, N. American Whites, S. African Blacks
		+	S. European, N. American Blacks
		±	Italy, Greece
	*0404	++	N. Europe, N. American Whites, S. African Blacks
		+	Chinese, Japanese, Polynesia
		±	Jews, Italy
	*0405	++	Japan, Polynesia
		+	Chinese
		±	Jews, Greeks, S. African Blacks, UK
HLA-DR1		+	S. Europe, Indians
		±	N. Europe, Jews, N. American Whites
HLA-DR10		++	Spanish, Jews
		+	Greeks, N. American Indians
		±	UK
HLA-DR6		++	Yakima American Indians

Table 1.2.1 adapted from Wordsworth and Brown [Wordsworth, 1998] summarises the known of HLA DR associations with RA amongst different ethnic groups. The strength of the association is illustrated by the use of a + or – sign, ++ indicating a strong and consistent association, + a weaker but consistent association and ± a weak or variable association

1998). Therefore, it is the DNA sequence between codons 67 to 74 rather than any particular HLA allele which is the principle determinant of genetic susceptibility to RA.

HLA-DRB1*04 and DRB1*01 genes are fairly common within the Northern European Caucasian population (population frequency of 30 – 40% for DRB1*04 and 5 –15% for DRB1*01) but most individuals with this genotype are unaffected by RA. The risk ratio for RA is 1 in 20 for the genotype DRB1*0404 / DRX, 1 in 35 with DRB1*0401 / DRX and 1 in 80 for DRB1*0101 / DRX (assuming a disease frequency of 1%). The absolute risk for a compound heterozygote with the genotype DRB1*0401 and DRB1*0404 is approximately 1 in 7 and some studies suggests this higher gene dosage also correlates with severity (Nepom and Nepom, 1998). The high population frequency of DR4 and DR1 makes screening for RA by HLA genotyping impractical and although there are advocates for its use in conjunction with testing for rheumatoid factor as a prognostic marker the evidence is inconsistent (Drossaers-Bakker *et al.*, 2002; Eberhardt *et al.*, 1996; Gough *et al.*, 1994; Harrison *et al.*, 1999).

1.2.4 Complex inheritance

1.2.4.1 Threshold model

Complex inheritance can be fairly well explained by the threshold model. This suggests that a complex disease, such as RA, is a result of an individual inheriting a certain number of risk alleles from a range of potential genes. The number of 'risky' alleles inherited must reach a *threshold* level of genetic risk before an individual can be described as having a genetic susceptibility to the disease. The individual must then encounter or be exposed to a *threshold* of environmental factors or triggers. The combination of reaching

both the genetic and environmental thresholds results in the development of the disease (Horwitz, 2000; Sham, 1998).

The recurrence risk for a complex disease is proportional to the population risk of the phenotype (section 1.2.2.2). The more 'disease genes' in the population, the greater the risk of disease recurrence in the offspring. The risk is also greater if more than one family member is affected as there will be more predisposing genes within the family and consequently the greater the risk of transmission from either or both parent(s).

This threshold for developing a complex disease may vary between men and women as is true of RA where the mean incidence rate is 3:1 in favour of women. The ratio is higher at a young age of onset implying that a man who develops RA under the age of 45 must have many more risk factors than a women developing disease at the same age, because the overall disease threshold is higher in a man. As a consequence the recurrence risk is greater in the offspring if the proband is male (Horwitz, 2000).

1.2.4.2 Models for genetic interaction

A multifactorial disease will generally involve two or more genetic loci that may interact in several ways. Epistasis is the term used to describe the interaction of two or more loci. An additive model describes the situation where the penetrance (section 1.2.4.3) of a disease is represented by the *sum* of the separate terms contributed by two or more loci. In a multiplicative model, the penetrance of a disease is represented by the *product* of separate terms (or penetrance factors) contributed by two or more loci. The heterogeneity model, assumes that genetic loci act *independently* (i.e. not epistatically) so that an individual can become affected through possessing a particular genotype at a locus regardless of the alleles at the other loci (Sham, 1998; Terwilliger and Goring, 2000).

The number of genes involved in the predisposition to a complex disease and whether each of the contributory genes have equal weight in conferring risk are other factors to be determined when identifying a model. A polygenic model implies an infinite number of genes involved, all contributing a small but equal risk of disease. An oligogenic model suggests fewer genes (maybe between 5 to 10). Some of these genes may have major effects in conferring risk whereas others might make relatively minor contributions.

Identifying the genetic model underlying a complex disease such as RA is important because knowing how the loci interact allows certain assumptions to be made during the analysis which can improve the power of a linkage study (section 1.2.8.2). It is also helpful when attempting to show exclusion of a particular chromosomal region (section 1.2.8.3). However, it can be difficult and if incorrectly identified can lead to power reductions (section 1.2.8.2). A way of choosing an appropriate multilocus model is to examine the recurrence risk in first, second and third degree relatives. The relative risks will decrease as the relationship becomes more distant. The rate of decrease can suggest a model for the disease loci interactions (Risch, 1990b; Risch, 1990c). For instance, under the additive model (which also applies for monogenic traits) the risks to successive degrees of relatives falls off at a steady rate but under the multiplicative model the recurrence risks decrease at a faster rate. The recurrence risk for relatives of an RA proband illustrated by fig 1.2.1. The decrease in relative risk in second and third degree relatives fits well with a multiplicative model and the rapid fall off in disease risk is clearly seen.

Parametric linkage analysis assumes that certain factors are known including the mode of inheritance, the allele frequencies, the phenocopy rate, the order and position of all markers, that all the diagnoses are correct and that the affectation status is known. Non-parametric linkage analysis (NPL) is used

where no clear model of inheritance can be identified. The model assumes there is no interaction between alleles at different loci. If the correct model is known parametric linkage analysis is more powerful than NPL quite but significant power is lost if an incorrect model is used. Hence, NPL is often used when studying complex multifactorial diseases.

1.2.4.3 Penetrance and phenocopies

The penetrance of a particular genotype at a disease locus is the probability that individuals possessing that genotype will develop that disease. The penetrance of a disease is complete if all individuals with the high-risk genotype will eventually develop the disorder (if they live for long enough). The penetrance of a disease is incomplete, reduced or partial if not all individuals possessing the disease genotype develop the trait. Incomplete pentrances may be a function of many variables including age, sex, environment and the presence of certain alleles at other genetic loci.

An individual is said to be a phenocopy if they manifest the same phenotype (or disease) as individuals having a particular genotype but do not possess the high-risk genotype themselves. It is a term often used to refer to a disease state caused by non-genetic effects. The power of any method to detect linkage is reduced by phenocopies in a data set. The phenocopy rate is defined as the proportion of cases in the population that are phenocopies, while the sporadic risk is defined as the probability of the disorder developing in individuals without a high risk genotype (Sham, 1998). The effect of any particular disease locus will be partially obscured by phenocopies due to other disease loci or to environmental effects, an inaccurate clinical diagnosis or random errors in data gathering and transcription. The problem of an inaccurate clinical diagnosis is increased if the clinical criteria used to define a disease are too broad. This is a continuing problem with a disease such as RA or a connective tissue disease.

1.2.5 Meiosis

All human cells are ultimately derived from a single cell, the zygote, formed by the union of two gametes, the ovum and the sperm. Each gamete contributes a half (or haploid) set of 23 chromosomes to form a full (or diploid) set of 23 pairs of chromosomes in the zygote. Normal gametes contain 22 autosomes and one sex chromosome. Two chromosomes are homologous if they belong to the same family of chromosomes e.g. both chromosome 16. Homologous chromosomes are similar in length and sequence and as such humans have two copies of every gene. Every time a cell division occurs in the zygote the 23 pairs of chromosomes are duplicated and this is called *mitosis.* Gametes however undergo *meiosis* where daughter cells containing only a haploid set of 22 autosomes and a sex chromosome are formed. This ensures that the union of two gametes will produce the correct number of 23 chromosome pairs. A recombination event occurs during meiosis when two chromosomes meet at a chiasma (or crossing point) and then separate leading to an exchange of genetic material between homologous chromosomes. Chiasma may occur at approximately equal frequency at any point along the chromosome. Chromosomes consisting of alternating segments of paternally and maternally derived DNA are formed. The reconstituted chromosomes then segregate independently into two daughter cells and these develop into gametes. With each generation the length of the DNA segments remaining in continued proximity to each other becomes shorter.

1.2.5.1 Mendel's laws of independent segregation

Mendel's first law of independent segregation states that genes occur in pairs and during gamete formation one of each pair is passed down to each gamete. The second law states that Genes controlling different characters segregate independently (Jacquard, 1978). This law is violated when two loci

are linked, i.e., when the recombination fraction between them is less than 0.5 (section 1.2.7.2)

1.2.5.2 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium or law states that the frequency of alleles in a large population will be constant from one generation to the next, providing mating is random (i.e. the genotype does not influence mating) and the genotype has no selective effect on the success of the production of offspring. There are many possible explanations for a significant deviation from Hardy-Weinberg equilibrium. One possible reason is that the genetic basis of the trait has been mis-specified. Another reason is non-random mating in the population. One form of non-random mating is population stratification where matings between individuals from different strata are less likely to occur than matings between individuals from the same stratum. Another form of non- random mating is assortative mating, where the probability of mating between two individuals is related to their phenotype similarity. Non-random mating may also arise from prohibition against or encouragement for matings between certain classes of relatives. Even when matings are random deviations from Hardy-Weinberg equilibrium can arise. Differential survival distorts the relative frequencies of possible mating types and consequently the distribution of genotypes among offspring. The preferential loss of individuals with certain genotypes can have a direct effect on the distribution of genotypes in the surviving population. Preferential selection for individuals with certain genotypes in the sampling process can also cause deviation from Hardy-Weinberg equilibrium (Horwitz, 2000; Sham, 1998).

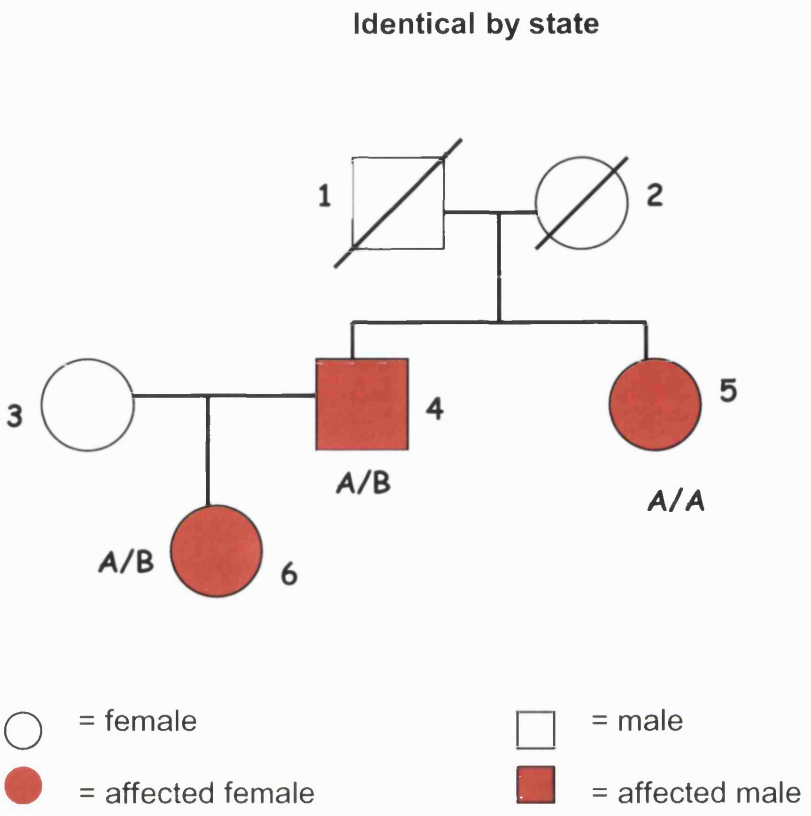


Fig 1.2.4 illustrates an extended family where individuals 4, 5 and 6 have RA (as indicated by symbols with a red colouring). Male or female gender are indicated by either a square or a circle respectively and a line crossing diagonally through a symbol denotes that the particular individual is dead. This applies to individuals 1 and 2 who were the parents of individuals 4 and 5. Hence, it is impossible to tell definitively whether the 'A' allele shared by individuals 4 and 5 comes from the same parental chromosome because the parental information is completely missing. The 'A' allele is therefore described as identical-by-state (ibs) i.e. although the same marker is the same it may not have come from the same parental chromosome. Without the genotype of individual 3 it is impossible to determine whether individual 6 shares the A or the B allele in common with her father. With extremely polymorphic loci the ibs relationships will approximate the ibd relations

1.2.5.3 Identity by state and identity by descent

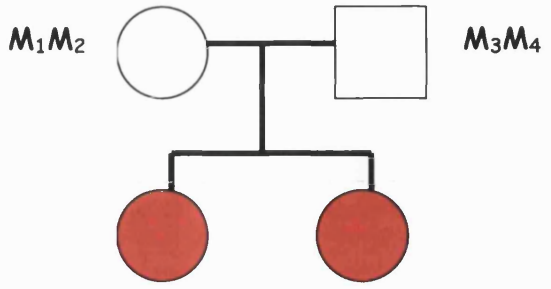
The aim of linkage and association analysis is to identify excess sharing of certain alleles or haplotypes either in related individuals (linkage analysis) or unrelated people who share the same disease phenotype (association analysis). Linkage analysis can be undertaken using affected sibling pairs (ASP) who have been diagnosed with the same condition. The null hypothesis is that there is no excess sharing of specific alleles. However an allele can be identical by state (ibs) or by descent (ibd). Alleles may have the same value but they may not have descended from the same specific ancestral allele. In these circumstances they are said to be identical by state (fig 1.2.4). Testing for linkage based on ibs sharing is heavily dependent on the estimates of the allele frequencies used because the expected parental alleles are based on those estimates.

Two alleles are said to be identical by descent (ibd) if both alleles are descended from the same specific ancestral allele (fig 1.2.5). It is possible for two individuals to have the same allele without the alleles being ibd, e.g. children of a homozygous parent who each inherit a different parental chromosome.

1.2.6 Identifying disease susceptibility genes: linkage and association analysis.

The classic strategy used to identify disease-causing genes begins with linkage analysis (fig 1.2.6). Systematic linkage screening of the entire human genome has the potential to detect all disease susceptibility loci. A methodical process of establishing linkage initially by coarse and then by fine mapping is undertaken. Once a region has been defined all genes mapping to the area are studied and mutations identified. Association studies can then be undertaken to evaluate biologically plausible candidate genes. The

Identical by descent



ibd	probabilities
Zero alleles shared	1/4
One allele shared	1/2
Two alleles shared	1/4

Fig 1.2.5 illustrates a nuclear family where both children have RA (as indicated by symbols with a red colouring). Male and female gender are indicated by a square or a circle respectively.
Individuals are identical by descent (ibd) if they share the same alleles and the ancestral origin of those alleles is the same. This can only by determined if both parents are heterozygous for the marker allele and all parental genotypes are available (unless sufficient unaffected siblings can be genotyped to infer the missing parental genotype). Power is improved by using ibd analysis as error rates are reduced (section 1.2.7.5).
If there is no linkage, the probabilities of the ASPs sharing zero, one or two alleles in common is 1/4, 1/2 and 1/4.

alternative approach used to identify disease causing genes is to perform 'candidate gene' studies using association analysis alone to evaluate a number of candidate genes without first performing a linkage study. Both methods are often used simultaneously (Terwilliger and Goring, 2000) (fig 1.2.6).

Linkage analysis is methodical but time consuming and expensive, both financially and in terms of DNA expenditure. It essentially generates hypotheses by identifying chromosomal regions where there may be candidate genes. The candidate gene approach is simpler, less costly but likely to miss a large number of potential candidate genes if used alone because only 15 – 20% of the 35,000 human genes have been identified to date. Association analysis is more powerful than linkage analysis but only over small chromosomal regions (i.e. thousands rather than millions of base pairs). Figure 1.2.13 illustrates the type of family that need to be recruited for linkage and association analysis.

1.2.7 Linkage analysis

1.2.7.1 Linkage analysis: the past, present and the future.

Linkage analyses using markers to follow disease inheritance have been undertaken since the 1950s in various different guises. Initially the markers were biochemical, based on blood groups and HLA, when at the time the chromosomal location of these markers was unknown.

In the 1980s restriction fragment length polymorphisms (RFLPs) were developed, based on the premise that for some marker sequences adjacent to the gene of interest, there will be inter-individual differences in the length of

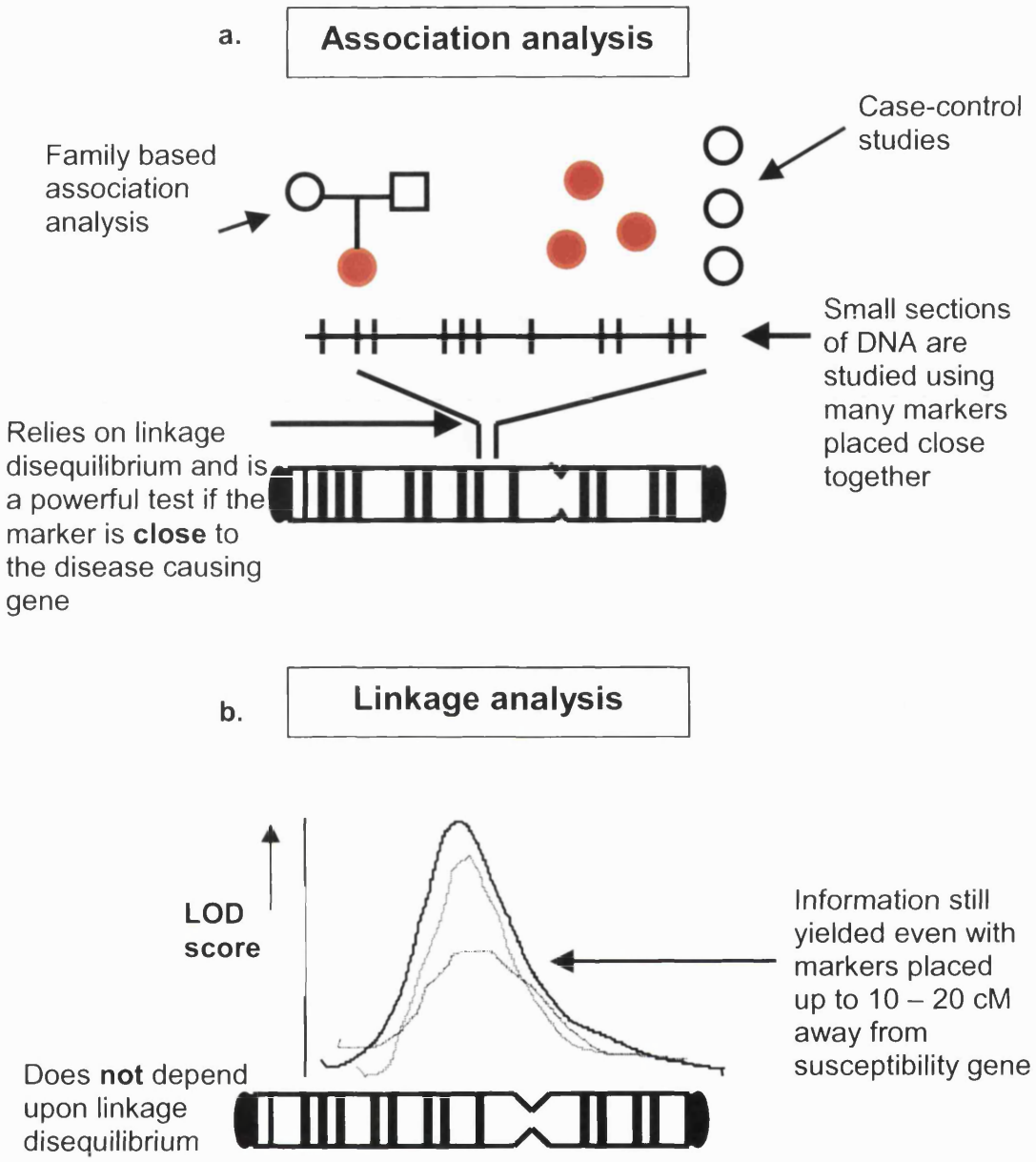


Fig 1.2.6 Illustrates the significance of and differences between association and linkage analysis.

Fig 1.2.6a shows that association analysis can be performed using either a case control or a family based study design. Association analysis relies on linkage disequilibrium and therefore markers need to be placed close to the disease causing gene for the test to have the required power. For susceptibilty loci with a low relative risk (1.5 – 4) association methods can provide substantially more power than nonparametric linkage studies such as sibling pair analysis.

Fig 1.2.6b shows that linkage analysis does not depend upon linkage disequilibrium. No assumptions need to be made regarding mode of inheritance and information is still yielded even when markers are placed 10 – 20 -cM away from the disease causing gene. A marker with a high LOD score ($\log_{10}$ of the odds for linkage) suggests that the disease causing gene may be nearby and is of greater significance if the linkage peak is narrow.

the sequence between two restriction cleavage sites. However, the technique was cumbersome, DNA usage high and the markers not very informative. For a marker to be informative each parent should be heterozygous and genotypically distinct from her mate to allow inheritance to be easily traced (section 1.2.5.3). Hence for a marker to be informative a dozen or more alleles will be identified in the general population. Most RFLPs have just two or three different alleles but microsatellite markers (section 1.2.7) derived from polymorphic repetitive sequences of DNA, are highly informative.

Most linkage studies (to date) have been performed with PCR-based microsatellite markers consisting of variable numbers of dinucleotide CA repeats or tetranucleotide repeats of other sequences. Microsatellite technology is faster than RFLP, the markers are informative and less DNA is used. More than 5,000 microsatellites have been described widely dispersed throughout the genome (fig 1.2.7). In the future, most linkage studies will be undertaken using a third generation of markers called single nucleotide polymorphisms (SNPs). These have the advantage of high throughput, high marker frequency as SNPs can be found approximately every thousand base-pairs and genotyping is undertaken using new hybridisation technology [DNA chips]. Their disadvantage is that there can be four alleles in the population for any given SNP (i.e. at any single position there will be either A, C, T or G). But with the combination of high marker frequency and throughput, although the information content (section 1.2.8.2) per marker will be less, the overall information will be far greater if SNP haplotypes are studied. This has the potential to revolutionise the genetic analysis of complex traits by studying large numbers of people with sufficient power to examine genes with weaker effects.

Microsatellite markers and PCR

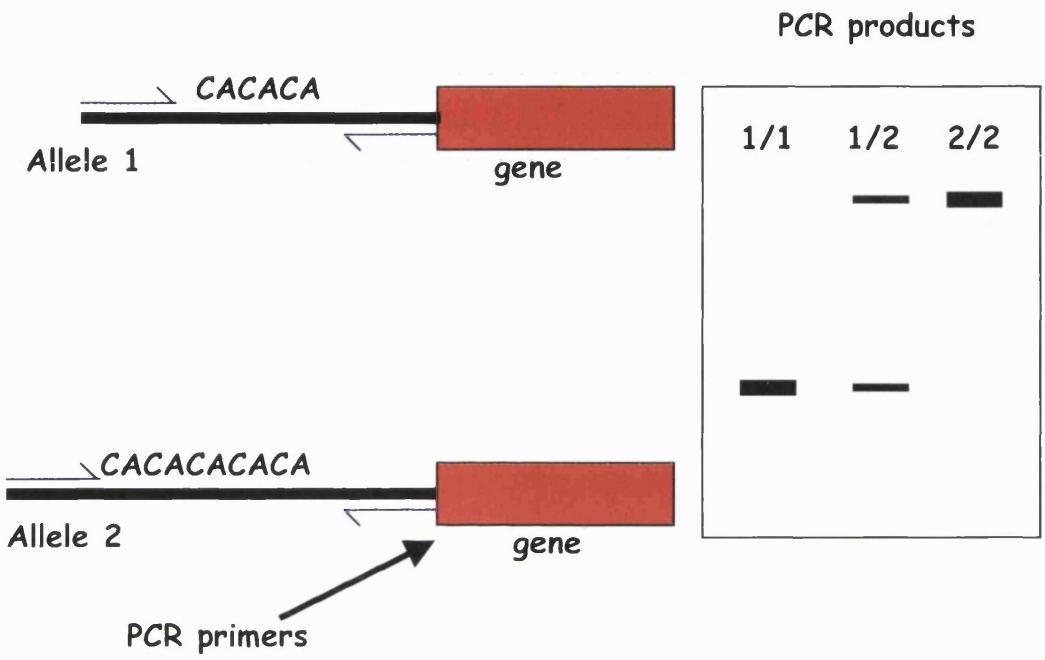


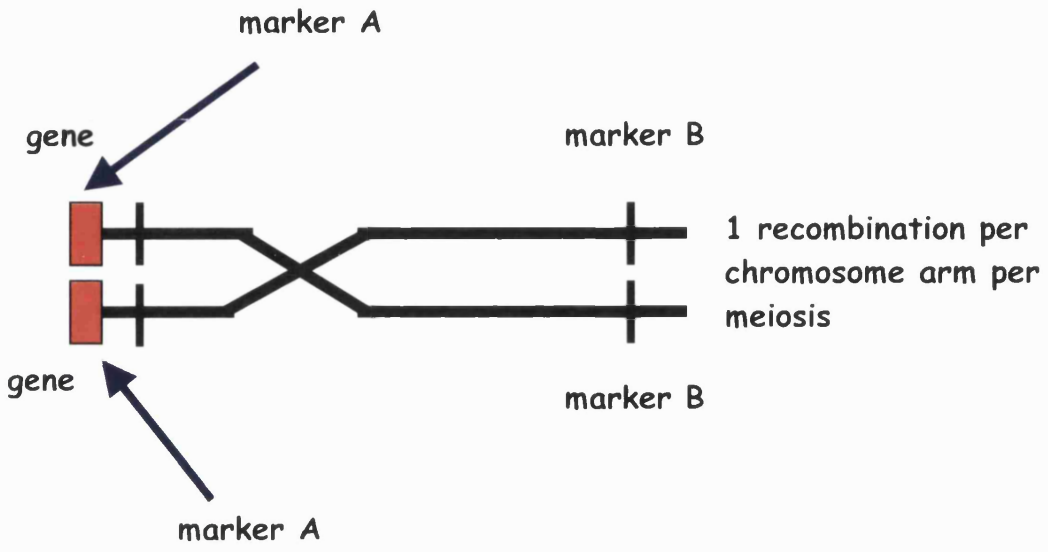
Fig 1.2.7 illustrates a microsatellite situated close to a gene and shows two of the microsatellite alleles. The PCR primers are designed to be complementary to short DNA sequences (approximately 20 bases) either side of the microsatellite (itself a series of dinucleotide CA repeats). The section of DNA between the primers is then amplified by PCR. In this example only alleles 1 and 2 of the microsatellite are shown although most microsatellites used in linkage analysis have between 10 to 20 alleles allowing for greater heterozygosity. Once the microsatellite is amplified the resulting products are separated using gel electrophoresis. Shorter fragments travel further and therefore allele 1 fragments will separate from allele 2 fragments and two bands will be seen if an individual is heterozygous for alleles 1 and 2.

1.2.7.2 Linkage analysis: recombination fraction

The goal of linkage analysis is to determine whether two loci segregate independently during meiosis. Approximately one recombination event per chromosome occurs during meiosis, so the alleles of loci on different chromosomes segregate independently of each other. The alleles of loci on opposite ends of the same chromosome also segregate independently. However, when two loci are close together on a chromosome their alleles will be co-inherited more than 50% of the time and such loci are said to be *linked* (Terwilliger and Goring, 2000) (fig 1.2.8). The closer loci are to each other on a chromosome, the lower the probability of recombination of their alleles. This probability is referred to as the *recombination fraction* θ. Alleles at two linked loci tend to be co-inherited because any cross-over of genetic material during a given meiosis is unlikely to occur because of the physical proximity. Hence, θ is correlated with (or proportional to) the physical distance between loci and the tightness of linkage is defined by recombination frequency (Horwitz, 2000; Terwilliger, 2000). The recombination fraction, θ, ranges between 0 (completely linked) and 1/2 (unlinked) and the essence of linkage analysis is to estimate the recombination fraction θ and to test whether $\theta = 0.5$ (null hypothesis).

The *genetic distance* between two loci is measured in terms of the expected number of cross-overs of genetic material, between loci, each meiosis. One Morgan (M) corresponds to the genetic distance over which one cross-over per meiosis is expected. One centiMorgan (cM) is equivalent to a 1% probability of a recombination per meiosis. As genetic distance is generally proportional to the physical distance between two loci, on average 1-cM corresponds to about 1 one million base pairs (one megabase (Mb)) of DNA (Horwitz, 2000). However, the correlation appears to be vary throughout the genome (Watkins *et al.*, 1994). Genetic distance and recombination fraction in males and females between the same markers may be different, with female values tending to be higher. When the position of one locus in the

Meiosis and recombination



Co-segregation likely with the nearest marker A but less so with marker B

Fig 1.2.8 illustrates the likelihood of recombination between a gene and a marker during meiosis and depends upon the genetic distance between the marker and gene. One recombination per chromosome arm is expected during each meiosis. Marker A is more likely to remain co-segregated with the gene than marker B because it is nearer the gene. Linkage analysis tests for excess sharing of marker alleles between affected individuals, if recombination occurs between the marker and disease causing allele any excess sharing will not be identified. Hence markers where excess sharing has occurred are likely to be close to a disease susceptibility gene.

genome is known, genetic linkage can be used to estimate of the chromosomal position of a second locus relative to the first. Genes predisposing to a phenotype can therefore be mapped relative to a large number of marker loci located at known positions throughout the genome. If these markers cover the whole genome this is described as a genome wide linkage analysis or whole genome screen. It is important to remember that linkage refers to the co-inheritance of two alleles at two loci that are physically linked on the same chromosome. It has nothing to do with phenotypic effects and a marker locus or genotype cannot be linked to a (disease) phenotype.

1.2.7.3 Assumptions made in linkage analysis

There are many assumptions made in linkage analysis. Some apply equally to parametric and nonparametric linkage analysis; other assumptions apply solely to parametric linkage analysis. As discussed above (section 1.2.4.2) parametric analysis assumes that certain factors are known including the mode of inheritance, allele frequencies, phenocopy rate and the order and position of all markers. Non-parametric analysis is frequently used where no clear model of inheritance can be identified and this model assumes there is no epistasis between alleles at different loci. It is also assumed that Mendelian segregation occurs with allele transmission and it is often assumed that there is no chiasma interference. Recombination between the marker locus and the disease locus is assumed not to have occurred but this cannot be the case in all circumstances. Any potentially common environmental effects are ignored and it is assumed that the phenotype depends solely on the genotype (penetrance 1.2.4.3). Random or non-assortative mating is assumed to have occurred and that the alleles of the study population are assumed to be in Hardy-Weinberg equilibrium. Finally, it is assumed that there are no genotype classification errors present in the data. This cannot be avoided in practice and occasionally the recombination status of a gamete may be misclassified (Terwilliger and Goring, 2000). This

is one of the reasons why some analysis methods are not as robust when used for analysis of real data, despite performing well on simulated error-free data.

1.2.7.4 Linkage analysis using affected sibling pairs

The original sibling pair method was proposed by Penrose in 1935 (Sham, 1998; Terwilliger and Goring, 2000) and is based upon the premise that under normal circumstances, there is an equal chance of two siblings inheriting the same parental allele at any particular marker locus. Therefore the probability that they share the same marker allele is 1/2. However, if both siblings are affected by the same genetically based condition they may be more likely to inherit a common allele from one or both parents (i.e. probability > 1/2). The chromosomal region surrounding the common disease allele should also be inherited by both. Therefore if a large number of affected sibling pairs (ASPs) were collected and genotyped the number of times a particular allele was inherited by both ASPs could be counted. The null hypothesis is that allele sharing between siblings is purely random. However, if there is excess sharing of a marker allele this implies linkage (figs 1.2.8 and 1.2.9). Affected sibling pair analysis does not depend upon a particular disease model and is often called non-parametric or model-free linkage analysis.

Non-parametric linkage analysis is most powerful when the identity-by-descent (ibd) or ancestral origin of the marker alleles can be determined (section 1.2.5.3). This is because ASP analysis tests whether each affected sibling pair shares 0, 1 or 2 alleles identical by descent. Therefore, the availability of parental DNA for marker-typing and highly informative markers is important. For a qualitative trait, ASPs should share alleles identical by descent more often than expected under random Mendelian segregation. For a quantitative trait, sibling pairs should show a correlation between the

Linkage

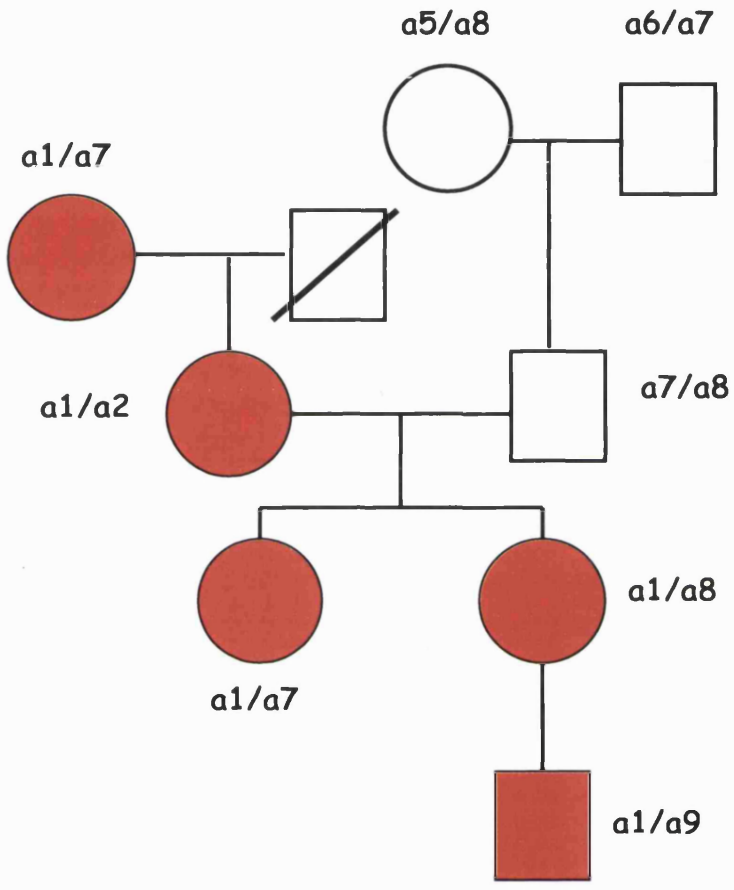


Fig 1.2.9 illustrates an extended family covering four generations where a grandmother, her daughter, both grand-daughters and great grand son have RA (as indicated by symbols with a red colouring). Male and female gender are indicated by a square or a circle respectively and a line crossing diagonally through a symbol denotes that the particular individual is dead.
In this family the "a1" allele has been inherited by all individuals with RA. The a1 marker allele and the disease susceptibility gene are linked.

magnitude of their phenotypic difference and the number of alleles shared ibd.

This method only indicates deviation from random inheritance. It has a number of advantages over traditional linkage analysis as it is generally easier to collect nuclear families rather than extended families and siblings are frequently more closely matched for age and environment than other relative pairs. No prior assumptions about the disease model, including parameters such as mode of inheritance, penetrance, phenocopy rate or disease allele frequency, are required. Hence, as discussed earlier, this form of analysis is more robust because linkage will not be missed as a result of any incorrect assumptions (fig 1.2.10). However, the traditional parametric analysis under a correct model is more powerful and will provide an estimate of the recombination fraction.

1.2.7.5 Linkage analysis: study design.

The power of a sibling pair study to demonstrate linkage at high levels of significance falls dramatically when λ_s is less than 3.0 (Risch, 1990c). The number of pairs required to detect linkage at a given level of significance is approximately inversely proportional to the square of the λ_s (Sawcer *et al.*, 1997). An optimal study design for a whole genome screen using ASPs is therefore the staged approach (Holmans and Craddock, 1997). The key to this method is the balance between the power required to detect weak genetic effects and the false positive rate. A moderate sized WGS has limited power to detect a logarithm of the odds (LOD) score of 3.0 to genes with a $\lambda_s < 2.0$ but it has high (>90%) power to detect linkage at much lower significance levels, e.g. LOD score of 0.7 (nominal linkage, $p<0.05$). Many of the LOD scores (section 1.2.8.3) that exceed this threshold will be false positives, reflecting random excess sharing in the data set, but a proportion

Affected sibling pair analysis

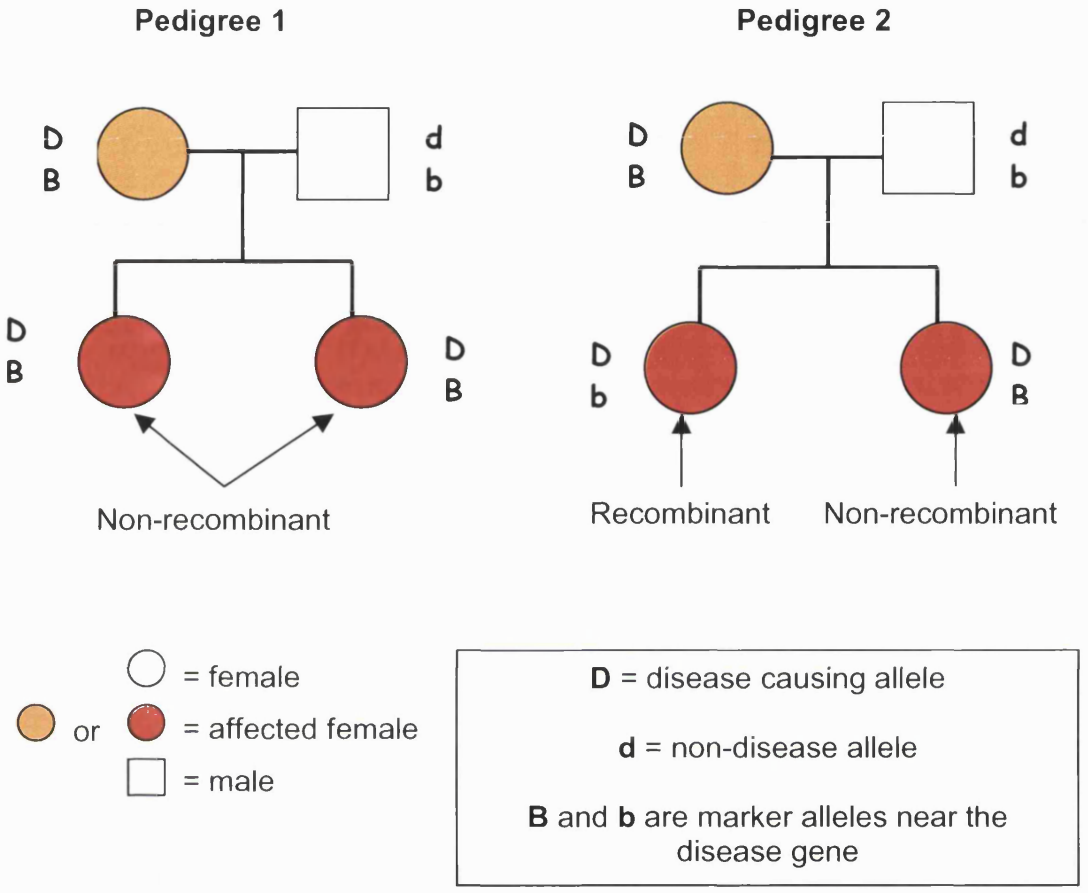


Fig 1.2.10 illustrates affected sibling pair (ASP) linkage analysis.
Pedigrees 1 and 2 represent nuclear families (i.e. mother, father and two affected siblings) where the mother and both offspring have RA. Marker B is close to the disease causing allele D on the mother's chromosomes. The father has alleles d and b and is not affected by RA.
Pedigree 1 illustrates the situation where both offspring have inherited the marker B and the disease causing allele D from their mother and no recombination has occurred between disease allele D and marker B.
Pedigree 2 however illustrates the situation where one of the offspring has inherited the disease causing allele D but not the marker B because recombination has occurred. The other offspring has inherited D and B as usual. Hence these ASP will not both have inherited marker B which in the mother was inherited along with the disease allele D. Linkage analysis tests for excess sharing between ASPs if recombination occurs between the marker and disease causing allele this will not be identified.

will reflect genuine linkage to genes will smaller effects. It is vital that the second stage has sufficient power to detect linkage at high levels of significance (LOD score ≥ 3.0) in order to distinguish reliably the true and false positives identified in the first stage.

Initially a sample large enough to demonstrate any weak effects (with a low significance) is studied. Marker density is low with markers spaced evenly over the genome. For a complex disease a map density with a mean distance of 10-20cM probably gives the optimal balance between effort and power. Additional families and markers are then typed in regions showing nominal evidence of linkage. Each locus (e.g. $p<0.05$) must be investigated thoroughly with more markers and the addition of more ASPs until a higher threshold is reached (LOD≥ 3.0, $p \leq 2 \times 10^{-5}$). If the size of the first data set is inadequate a weak effect may be missed and replication never attempted but if the size of the second sample is inadequate a true positive may not be replicated. Three factors predominately influence the power to detect a true positive. The first is the sample size; the second is the genetic effect and the third relates to the marker including the marker informativity and the distance from the disease locus. Genetic interactions, heterogeneity, penetrance and the frequency of a disease allele all affect the power of a study. Parental genotypes are valuable when checking for data error and can be used to determine a true allele frequency thereby avoiding the use of estimates (section 1.2.5.3).

The power of a study is frequently over-estimated for a number of reasons. Firstly, assumptions are made in the calculations that are frequently incorrect (section 1.2.7.3). In particular, it is assumed that the marker and susceptibility locus is completely linked and that the marker is fully informative (section 1.2.8.1). Although highly polymorphic microsatellite markers and multipoint analysis reduce the effects of recombination and non-informativity these assumptions remains at odds with reality. Secondly, error rates in the

diagnosis, genotyping and data transcription all have significant effects on power. Thirdly, the number of markers typed will alter the power but the relationship between marker numbers and power is not linear. For instance, a ten fold increase in genotyping is required to reduce the mean marker interval from 20cM to 2cM but this only increases the linkage information by two fold (Hauser *et al.*, 1996).

Variations in gene sharing between data sets are common. Genetic heterogeneity could explain these variations but random chance is probably responsible for most variations. In smaller sets of 100 or so fully informative sibling pairs values for ibd may vary dramatically. Data sets of 200 or more fully informative sibling pairs appear to provide more stable ibd values. A high density map of highly polymorphic markers is important when attempting to confirm real linkage as linkage evidence may decrease very quickly as the distance from the disease gene increases (Luo *et al.*, 1996).

1.2.7.6 Linkage analysis: avoiding bias

There are several potential sources for bias in linkage studies. For instance, missing parental genotypes can be most easily deduced when their children show a variety of alleles. This will occur most often in cases when the children do not share alleles ibd. Hence, deducing such parents and using the families may incorrectly bias the results away from sharing.

Some families have multiple affected offspring but these pairs will not all be independent of each other. A variety of weighting functions have been proposed to correct for non-independence in large sibships (Cordell *et al*, 2000; Greenwood *et al*, 1999; Hodge *et al*, 1984; Sham *et al*, 1997). One method is to use all possible pairs but correct for the fact they are not independent (MAPMAKER/SIBS). Other options are to select a single pair

from each family (but this is rather conservative) or to use a single affected sibling paired with all other possible affected siblings.

In an intercross mating both parents have the same heterozygous genotype and therefore in a number of cases it is impossible to distinguish whether ASPs share 2 or 0 alleles ibd. Although these cases may be omitted from the analysis other intercrosses where sharing can be determined may be included resulting in a false excess of sharing 1 allele. This would bias χ^2 statistic so intercross sibling pairs may be analysed with the t_2 statistic or with the MLS method.

Accurate assessment of the phenotype (i.e. diagnosis) is essential for linkage studies. Diagnostic criteria should be employed to standardise the information. To avoid bias the persons who establish the diagnosis should not be informed of genetic marker data and the laboratory personnel should not be informed of the diagnosis.

1.2.8 Approaches to data validation

Error detection is very important and needs to be considered at all stages of a genetic study. Errors include an incorrect diagnosis, mis-specifying genotypes, DNA mix ups, data transcription, non-paternity, mis-specification of a pedigree (e.g. identifying a half-sibling as a full sibling) etc. Errors are inevitable when dealing with hundreds of families and hundreds of microsatellite markers requiring the identification of 250,000 or more genotypes (section 1.3.3). By methodically increasing the level of testing, error checks can detect increasingly subtle errors and is the most successful way to eliminate genotype classification errors and avoid the misclassification of the recombination status of a gamete. Some error detection can be performed by manually checking all genotypes and pedigrees but higher

levels of checking require specific programs. Error detection can be broken down into five categories which include errors in the allele frequency used, Menedlian inconsistencies, strange segregation, haplotyping and relationship errors

Comparing the allele frequencies of study data with published allele frequencies (e.g. from Genethon or Marshfield databases) can allow for a number of factors to be considered. These include the comparison of the expected versus the observed number of homozygotes and heterozygotes. Data should fit with Hardy-Weinberg equilibrium and the microsatellite marker controls should correspond to the expected number of base pair repeats. Both water lanes and microsatellite marker controls should be run with all gels to identify contamination.

Mendelian inconsistencies can be checked for at four levels of increasing sophistication. Level 1 checks for simple errors and includes the manual checking for pedigree inconsistencies (e.g. if the alleles of a child and a parent are incompatible or there are more than four alleles in a sibship). This level of checking can also be undertaken using the program GAS (Alan Young, unpublished). Level 2 checks for more subtle errors. A genotype elimination algorithm (performed automatically by the program pedcheck if no paternal DNA is available) can detect these more subtle errors (O'Connell and Weeks, 1998). Level 3 checking uses a critical-genotype algorithm and attempts to identify the possible error by "untyping" one individual at a time and applying a genotype elimination (pedcheck) to see if the inconsistency has been eliminated. Level 4 checking using an odds-ratio algorithm and for each critical genotype, the relative likelihood of a different alternative genotype is determined.

"Strange segregation" checking looks for a more than expected dissimilarity in ibd sharing among siblings (or affected relatives) i.e. the preferential

segregation of an allele from heterozygous parents (transmission distortion). This type of checking can be undertaken using the δ parameter of the program Gene Hunter plus (Cox *et al.*, 1999b; Kong and Cox, 1997). Haplotyping using multipoint analysis can be used to identify excess double (or higher) recombinants within a small region. A number of programs are available to conduct multipoint analysis e.g. MAPMAKER/SIBS (version 2) (Lander *et al.*, 1987), SimWalk (Sobel and Lange, 1996) and Gene Hunter plus (Cox *et al.*, 1999b; Kong and Cox, 1997). SimWalk is useful in that it will calculate the probability of the error likelihood, however the calculations are extremely time consuming when studying hundreds of ASP families and markers.

Relationship error checking is particularly important in sibling-pair analysis of a late onset disease, such as RA, where mistakes may go undetected since parental genotypes are commonly not known. This level of checking can be summarised by two questions: "Which relationship is most likely?" and "How likely is the observed ibs sharing, conditional on the assumed relationships?". The verification of relationships can be determined using one of two programs RELPAIR (Boehnke and Cox, 1997) or Relative (Goring and Ott, 1997) which calculate the likelihood of the marker data, conditional on different assumed relationships. RELPAIR also looks for relationships across families allowing DNA mix ups to be identified and can identify the likely observed ibs sharing conditional upon the assumed relationship (as can 'siberr' by Wagner & Ehm (Ehm and Wagner, 1998)).

1.2.8.1 Marker informativity

A polymorphism information content (PIC) is the fraction of matings in which a parent is expected to be fully informative. A parent is fully informative if it is always possible to determine which of their alleles are inherited by their children. This will depend in part upon the genotype of their mate as

information for linkage comes from doubly heterozygous people. Microsatellites are highly polymorphic markers and are therefore informative with a polymorphic information content (PIC) of approximately 0.8 for most microsatellite markers. Single nucleotide polymorphisms (SNP) will be less informative for linkage as at any one position there can only one of four possible nucleotides and many people will therefore be homozygous at a specific marker. However, SNPs occur every thousand or so base pairs and SNP haplotyping could be used to increase their informativity.

1.2.8.2 Power

Power is a test of the probability of rejecting the null hypothesis given an alternative hypothesis is true (i.e. the likelihood of a type 2 error). It is necessary to study very large numbers of affected sibling pair (ASP) families if adequate power to detect linkage to non-MHC genes is to be generated (Brown and Wordsworth, 1998). Recruitment is difficult in a late onset disease such as RA and may require international collaboration. Undertaking whole genome analysis on such large numbers of individuals is both expensive and time consuming.

1.2.8.3 LOD score

This is a log of the odds (LOD) in favour of linkage of a genetic marker to a phenotype as opposed to non-linkage (random segregation). It is the logarithm of the likelihood ratio. A LOD of 3 (interpreted as 1000:1 odds in favour of a conditional probability of linkage) is arbitrarily accepted standard for linkage for an autosomal disease (section 1.2.8.4). A LOD score of 3 corresponds to a p value of about 2.2×10^{-5}. A LOD of 3.6 is expected to occur by chance in 1 WGS in 20 (fig 1.2.10). LOD scores can be added across independent families.

1.2.8.4 Levels of significance for whole genome linkage analysis

Corrections for multiple testing are required to avoid excess error in whole genome linkage analysis. In 1995 Lander and Kruglyak proposed a set of criteria designed to determine levels of significance for WGS (Lander and Kruglyak, 1995). Their proposals were based upon the number of likely false positives per WGS and are listed below.

- Suggestive linkage – this is where statistical evidence of linkage is expected to occur at random (i.e. a false positive), once in one WGS ($p<7.4 \times 10^{-4}$, or LOD (or MLS) = 2.2).
- Significant linkage - this is where statistical evidence is expected to occur 0.05 times in a WGS ($p <2.2 \times 10^{-5}$, or LOD = 3.6) i.e. once in 20 WGS.
- Highly significant linkage - this is where statistical evidence is expected to occur 0.01 times in a whole genome screen ($p <3 \times 10^{-7}$ or LOD = 5.4).
- Confirmed linkage – this is where significant linkage has been documented by one study or a combination of initial studies and has been subsequently confirmed in a further sample with a nominal p value of 0.01

The above standards are very difficult to achieve for complex traits, particularly those for highly significant and confirmed linkages. Several possibilities may explain the failure to confirm linkage in these studies. Firstly, these linkages may be spurious and no disease gene is encoded in these regions. Secondly, it may take many more ASPs to confirm these genes because they have such weak influences on the disease phenotype. Finally, it would be difficult to obtain significant linkage in the overall data set if the linkage evidence is heterogeneous from one subset to another.

1.2.9 Various forms of analysis

1.2.9.1 Two point and multipoint analysis

Two-point (or single point) analysis fails to make use of the full inheritance information available as it is based on studying individual genetic markers one at a time i.e. if the number of shared alleles cannot be calculated unambiguously certain ASPs are not useful in the analysis. It is not always possible to determine the exact ibd status at every marker locus, particularly when parents are unavailable for study or are not fully informative. Under these circumstances only ibs information for the offspring can be determined. Equally, the ibd status at non-marker locations cannot be assessed.

Multipoint analysis uses the genotype information for each sibling pair, together with information about parents and additional siblings (where available) to infer the ibd distribution at each point along the genome. Information from nearby markers is used to increase the informativity at a potentially ambiguous marker (Fulker and Cardon, 1994). Multipoint analysis also allows for the construction of haplotypes which is important when checking for genotype errors.

MAPMAKER/SIBS is a statistical package that can be used to undertake both single and multipoint analysis. A number of other sibling pair analyses can be performed using MAPMAKER/SIBLINGS and these include:

- Exclusion mapping. This tests specific hypotheses about the degree of sharing at each location in the genome i.e. it is used to identify and exclude those regions unlikely to have a major effect on the trait and hence allow attention to be focused upon the remainder of the genome.
- Maximum likelihood mapping. This identifies loci involved in a qualitative trait.

- Information content mapping. This assesses the extent to which the available genetic markers have extracted the full inheritance information at each location in the genome.
- Quantitative trait locus (QTL) mapping. This identifies loci involved in a quantitative trait by two parametric methods and one non-parametric method.

MAPMAKER/SIBS uses the Lander-Green algorithm (Lander *et al.*, 1987) and as such is particularly suitable for small nuclear families or sibships since the amount of computation increases only linearly with an increase in the number of marker locations but exponentially with the size of the family.

The algorithm works by completely specifying, by an inheritance vector $V_i(S)$, the inheritance pattern of each pedigree P_i, at each location S (Lander *et al.*, 1987). Each component of this vector corresponds to a particular meiosis and the co-ordinate 0 or 1 is given according to whether the offspring inherits the allele, (at position S) from the paternally or maternally derived chromosome. For each ASP the vector therefore has four components - two for each sibling. One specifies whether the mother's maternally or paternally derived allele has been transmitted to the sibling and the other specifies the corresponding information for the father. The probability distribution for $V_i(S)$ at each location on the genome can be calculated conditional on the genetic marker data. To calculate whether the i th sibling pair shares 0, 1 or 2 alleles ibd at position s requires only the addition of the probabilities of the appropriate vectors.

For a qualitative trait, each locus can be characterised by the expected proportions Z_o, Z_1 and Z_2 of the affected sibling pairs sharing 0, 1 and 2 alleles ibd. Mendelian proportions (α_0, α_1 and α_2) are 1/4, 1/2 and 1/4. Holmans defined the "possible triangle" in 1993 and z_o, z_1 and z_2 will lie within this:

i.e. $Z_o + Z_1 + Z_2 = 1$

and $Z_1 \leq 1/2$ [1]

and $Z_1 \geq 2(Z_0)$

If no dominance variance is assumed, this is the equivalent to the constraint

$Z_1 = 1/2$ [2]

the sharing proportions are then described by the single parameter Z_2

If only a single locus is involved, the sharing proportions (Z_0, Z_1 and Z_2) can also be expressed in terms of relative risks (Risch 1990a). If λ_s is the relative risk ratio for a sibling, λ_o is the relative risk ratio for an offspring and λ_m is the relative risk ratio for a monozygotic twin, then the following relationships hold:

$$Z_o = \alpha_0 / \lambda_s;$$
$$Z_1 = (\alpha_1 \times \lambda_o) / \lambda_s; \quad [3]$$
$$Z_2 = (\alpha 2 \times \lambda_m) / \lambda_s;$$

λ_s, the relative risk ratio for a sibling, is defined as prevalence in the siblings of affected individuals divided by the population prevalence.

If no dominance variance is assumed:

$$\lambda_o = \lambda_s$$
$$\text{and} \quad \lambda_m - 1 = 2(\lambda_s - 1) \quad [4]$$

The relationships hold if multiple loci are involved in the trait and interact multiplicatively.

1.2.9.2 Maximum likelihood mapping

To identify regions of significant excess allele sharing, the maximum likelihood values of the allele-sharing proportions (Z_0, Z_1 and Z_2) are estimated at each location along the genome. The maximum LOD score Z(*S*) at each location is computed. The likelihood of the observed data (arising under these maximum likelihood values) is compared to the likelihood under random Mendelian segregation. MAPMAKER/SIBS allows the calculation of the maximum likelihood proportions (Z_0, Z_1 and Z_2) in two forms: subject only to the "possible triangle" constraint (equation 1) or subject to the additional constraint of no dominance variance (equation 2). A LOD score can never be

negative because the maximum likelihood solution (Z_0, Z_1 and Z_2) at each location is never worse than the random Mendelian segregation. To limit the genome-wide false-positive rate to 5%, a LOD score threshold of 4.0 is used when implementing the "possible triangle" method. This corresponds to a single-test significance level of 2×10^{-5} and the calculations are based upon the use of an infinitely dense map. However, this significance level is recommended even when using moderately spaced markers for two main reasons. Firstly, it is anticipated that the marker density will be increased in regions showing suggestive evidence of linkage and similar studies are likely to be carried out on the same sample. Secondly, the increase in stringency over and above the 5% level is comparatively small.

Information-content mapping makes clear whether the vast majority of the ibd information has been extracted. It focuses attention on the regions where additional markers would provide substantially more information e.g. increase or decrease the LOD score in a region with a suggestive result. A simulation study can be undertaken as part of the study design to explore the effects of map density, marker polymorphism rate and availability of parents for genotyping. Assuming a heterozygosity rate of 75% (similar to that of the ABI LMSvII marker set used by the UK WGS (chapter 2)), a 10-cM map extracts ~85% of the ibd information at the markers and approximately 70% midway between markers (when parents are available). This falls to 65% and 55% without parental DNA (Kruglyak and Lander, 1995). The increase in power of multipoint analysis over single-point analysis is greater in studies where parents are unavailable (Kruglyak and Lander, 1995). However, in these circumstances the marker allele frequencies are very important and if mis-specified can lead to false-positive results. Under-estimating the frequency of an allele will lead to over-estimating the degree of ibd. If reliable estimates of allele frequencies from an appropriate population are available they should be used. If this is not possible, a number of software programs such as RECODE or DOWNFREQ allow the estimation of allele frequencies to be

undertaken using all the unrelated individuals from the study or all the unrelated individuals and a proportion of the ASPs. A sensitivity analysis can be performed to test the sensitivity of the results to changes in allele frequencies. Also, the lower allele frequency boundary can be set for all alleles making it difficult to under-estimate allele frequencies and thus the corresponding LOD scores will be conservative.

1.2.9.3 Exclusion mapping

This is a method to exclude regions of the genome that are unlikely to encode susceptibility genes. A LOD score of –2 is the conventional threshold for excluding a site from linkage.

1.2.9.4 Genetic models and calculation of expected sharing

In order to determine the power of a linkage study or when attempting to show exclusion it is necessary to calculate the expected ibd allele sharing that would be observed at one of many possible disease loci. To do this it is necessary to model how the disease loci interact to determine the trait phenotype. A way of choosing an appropriate multilocus model is to examine the risks in relatives of differing degrees. The relative risks will decrease as the relationship becomes more distant. The rate of decrease can suggest a model for the disease loci interactions (Risch, 1990b; Risch, 1990c). For instance under the additive model (and this also applies for monogenic traits) the risks to successive degrees of relatives falls of at a steady rate but under the multiplicative model the recurrence risks decrease at a faster rate.

1.2.9.5 Conditional analysis

Conditional analyses of marker can be undertaken by dividing the total sample into subsets defined by variation at some other markers (e.g. HLA). Such stratification procedures have been used in a number of studies aiming to strengthen evidence for linkage. However, the interpretation of results can be difficult (Concannon *et al.*, 1998). For the procedure to enhance the power to detect linkage, the locus being tested must show a large difference in ibd amongst subsets. If the joint effects of two or more loci are non-additive, the differences seen in ibd after stratification will not be large and if the effects are exactly multiplicative no differences in ibd amongst the subsets will be identified. However, if the genetic effects are additive differences in ibd can be seen but may not be large enough to substantially increase the power to detect linkage.

Conditional analyses with modest sample sizes can produce inconsistent results as illustrated by the first WGS in insulin-dependent diabetes mellitus (Concannon *et al.*, 1998). Two groups reported that the evidence for linkage at FGF3 on 11q13 (IDDM4) was much stronger after they had conditioned on HLA sharing (or HLA type). However, one group found increased evidence for linkage in the subgroup of ASPs who shared HLA alleles but the other group found increased evidence in those ASPs who did not share at HLA (Davies *et al.*, 1994; Hashimoto *et al.*, 1994). Contradictory findings such as these cannot be interpreted as strengthening the evidence for linkage.

1.2.10 Association analysis

The essential difference between linkage and association is that linkage is a relationship between loci and association is a relationship between alleles. Even very close linkage is not sufficient to cause population association.

Population association between allele X and disease Y can occur for three reasons:

- Allele X can *directly cause susceptibility* to disease Y. If so, the same allele X should be associated with disease Y in any population studied, unless the cause of disease varies between populations.
- If the disease bearing chromosomes in the population are descended from a few ancestral chromosomes then very close linkage can produce allelic association at the population level. If *linkage disequilibrium* (section 1.2.10.1) is the cause of the association a gene should be discovered near locus X which has mutations in people with disease Y. However the particular allele at the locus X which is associated with disease Y may be different in different populations.
- Population stratification could be involved (sections 1.2.10.3 3.3.5). People with disease Y and those without may be genetically different subsets of the population who coincidentally may also have different frequencies of allele X. An example is the association between HLA-A1 and the ability to eat with chopsticks in the San Francisco Bay area (Lander and Schork, 1994). HLA-A1 is more frequent among Chinese than Caucasians.

Population based case-control studies of disease-marker associations can therefore be problematic for two main reasons. If suitable controls are not sought these studies may not be able to distinguish between linkage disequilibrium and population stratification. It is essential that patient and controls studied come from the same population so that any comparison of gene frequencies will be valid. Within family association studies e.g. transmission disequilibrium testing (TDT) have been developed to avoid some of these problems. TDT (Spielman *et al.*, 1993) can avoid population stratification but cannot distinguish associations caused by linkage disequilibrium from those where the marker itself is the susceptibility factor. These studies involve more work than the standard case-control study as

three individuals (proband and parents) are typed and in a late onset disease such as RA both parents may not be available. Secondly, if the statistical analysis is not rigorous enough inadequate corrections may be made for the number of questions posed. Each test performed carries an independent risk of a false positive result. A Bonferroni correction is sometimes applied and the threshold of significance is set at $p=0.05/n$ where n is the number of independent potential associations checked. If n loci with m alleles each are tested the rigorous correction factor would be $n(m-1)$. The inconsistency of published association studies may be due, in part, to this lack of rigorous correction.

Two 'unrelated' people in the UK would typically share common ancestors about 22 generations (500 years) ago and if fully outbred, would have $2^{22} = 4$ million ancestors each. In the 15th century the population of Britain was around 4 million. Assuming the UK population interbreeds freely, only about 44 meioses separate two unrelated individuals. Loci showing 1% recombination per meiosis would therefore have a better than 50% chance of remaining in the same combination through 44 meioses ($(0.99)^{44} = 0.64$). This suggests that allelic associations may be noticeable for loci within 1cM of each other. Given that the human genome is 2644cM a complete genome scan for markers in linkage disequilibrium with a disease would require at least 3000 markers. Corrections for multiple testing would need to be very robust suggesting only extremely strong associations would be significant after correction. Hence, testing for linkage disequilibrium is generally restricted to a candidate region where there is an a priori hypothesis.

1.2.10.1 Linkage disequilibrium

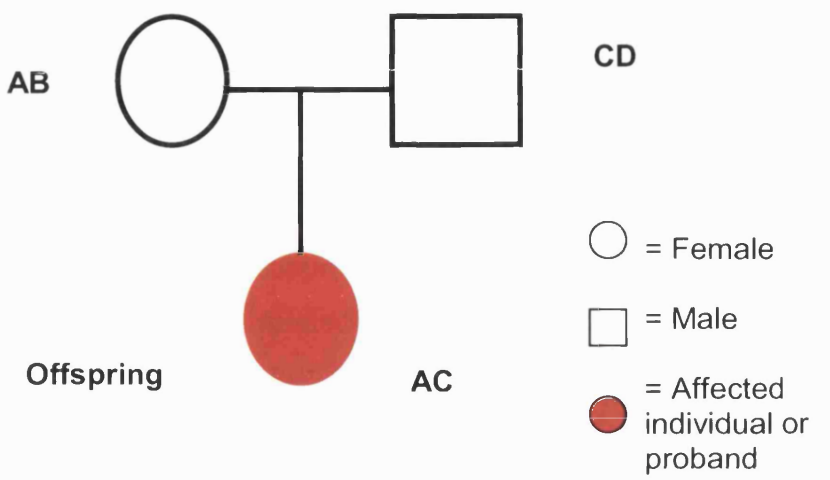
Linkage disequilibrium occurs when a particular marker allele is associated with the disease-trait locus at a greater than expected frequency across

multiple families. When a marker and a disease locus are very close together on a chromosome genetic crossover will have occurred at such a low rate that the marker will appear to cosegregate with the gene regardless of the family studied. This is in contrast to the situation where the two loci are further apart but still linked in which case repeated crossing over will allow all possible combinations of chromosomal haplotypes to appear with frequencies as predicted by the equation for Hardy-Weinberg equilibrium. In general, the stronger the disequilibrium the closer the marker to the disease locus. However, allele frequencies and mutation rates at the marker locus also effect the level of observed disequilibrium. Linkage disequilibrium may be due to recent population bottlenecks or new mutations. With time and further recombinations the nearby loci will return to an equilibrium.

The term linkage disequilibrium is sometimes used interchangeably with the term allelic association. Allelic association refers to a significantly increased or decreased frequency of occurrence of a marker trait in combination with a disease trait. It can be explained either by biological interaction of the marker allele with the disease-trait gene or by linkage disequilibrium.

The goal of linkage disequilibrium analysis is to map loci relative to each other and thereby estimate the genomic position of new loci of unknown position using loci of known location. Linkage disequilibrium mapping, although generally conducted using unrelated individuals can be viewed as a

Transmission disequilibrium testing (TDT analysis)



"Case" alleles = AC
"Control" alleles = BD

Fig 1.2.11 illustrates the principle of transmission disequilibrium testing or TDT analysis. This is a form of family based association testing.
The offspring (red circle) has inherited **A** and **C** alleles from the parents and the untransmitted parental alleles are **B** and **D**. TDT analysis compares the frequency of specific alleles transmitted to an affected offspring or proband with the frequency (in the parents) of the alleles that are not transmitted. The untransmitted parental alleles are used as the 'control' and the transmitted alleles as the 'case'.

linkage study involving an extremely large hypothetical pedigree with many generations of indeterminate structure and availability of DNA only in the bottom generations. As discussed above these individuals are in fact, distantly related. Association tests measure the linkage disequilibrium present between a trait allele and nearby marker loci.

1.2.10.2 Relationship between association and linkage analysis

Both linkage and association analysis have been designed to detect non-random co-inheritance of alleles at two loci. Linkage analysis depends on studying sets of related individuals. Association analysis appears to study unrelated individuals but it depends upon the fact that those subjects are, in reality, distantly related. Closely related individuals are descended from common ancestors via a small number of meioses and tend to have large segments of chromosomes in common. Distantly related individuals are separated by large numbers of meioses and therefore have only small chromosomal segments in common following numerous recombination events. This is the reason why association analysis has a finer resolving power compared to linkage analysis but also why it is not as sensitive to loose linkage. In some respects they represent two ends of a continuous spectrum. Association analysis could be seen as a form of identity-by-state analysis between very distant relatives. However, random events cannot be fully defined because the exact 'relationships' between the individuals cannot be fully defined.

1.2.10.3 Transmission disequilibrium testing (TDT analysis)

TDT analysis can identify both linkage and linkage disequilibrium. Only linkage disequilibrium can distort the distribution of marker genotypes among the parents of affected individuals in comparison to the distribution of genotypes in the general population. However, the ability to detect linkage disequilibrium only applies if the sample consists of unrelated cases and

parents (Sham and Curtis, 1995) but it is a valid test of linkage in all situations. TDT analysis considers the probabilities of marker allele transmission from heterozygote parents to affected individuals, and the distortion of these probabilities from 0.5 can only occur if the marker and disease loci are linked. This assumes there is no overall distortion from Mendelian segregation at the marker locus in the population. The test effectively considers the allele transmission from the father and mother of affected individuals to be independent events (fig 1.2.11). This assumption is valid if the recombination fraction is very small. Occasionally this is untrue and in those circumstances the TDT can lose some of the information available in the sample. However the assumption holds in most circumstances. The TDT can be used to test for association using both affected singleton data and affected sibling pair data. It was originally designed for bi-allelic markers but has been extended for multi-allelic markers (Spielman and Ewens, 1996) and quantitative traits (Abecasis *et al.*, 2000; Allison, 1997; Rabinowitz, 1997). It avoids the potential problems of population stratification that can occur in case control association studies because the 'control' alleles are the untransmitted parental alleles (sections 1.2.10 and 3.3.2). Possible sources of bias occur if parental DNA is unavailable when bi-allelic markers are used even if parental genotypes are inferred from the offspring genotypes (Curtis, 1997).

1.2.10.4 Monte Carlo Simulation

This is undertaken to provide an empirical estimate of the significance level of the maximum LOD score obtained from the actual data (fig 1.2.12) (Kaplan *et al.*, 1997). In each simulation disease phenotypic data are retained but marker phenotypic data are generated under the null hypothesis of non-linkage between the disease and marker loci. Therefore the pedigree structures and the disease phenotypes are kept but the marker phenotypes of

Monte Carlo Simulation

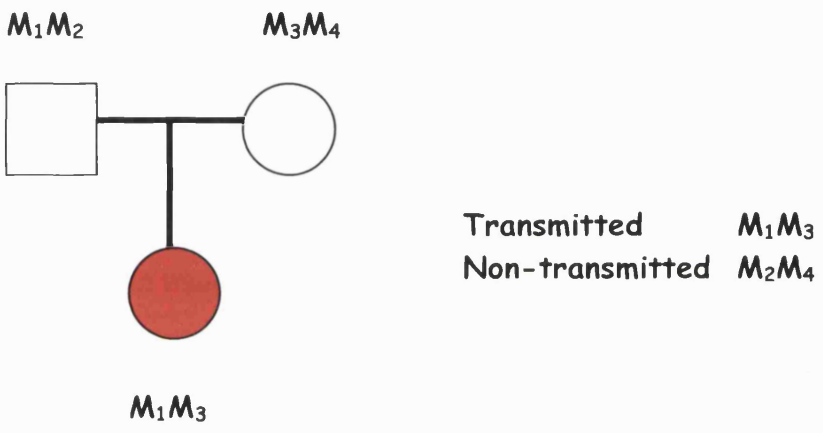


Figure 1.2.12 illustrates the principle behind Monte Carlo simulation. A series of simulated data sets are generated using the pedigree structure and disease phenotypes of the actual data set but with generated marker phenotypes based on population allele frequencies. The maximum LOD score is calculated and compared to the LOD score of the actual data. The proportion of simulated data sets for which the maximum LOD score exceeds the maximum LOD score of the actual data provides an estimate of the support for linkage.

the founders are generated according to the population allele frequencies and the marker phenotypes of the non-founders are generated according to Mendelian segregation without regard to disease phenotypes. For each simulated data set, the maximum LOD score is calculated and compared to the LOD score of the actual data. The proportion of simulated data sets for which the maximum LOD score exceeds the maximum LOD score of the actual data provides an estimate of the extent of support for linkage. Monte Carlo simulations also provide an estimate of how informative the data set is.

If there is no linkage the labels "transmitted and non-transmitted" can be permutated and the labels for each heterozygote parent in the sample are shuffled. The statistic is recomputed and then this procedure is repeated many times. The proportion of times that the statistic is larger than or equal to the observed value gives an estimate of p-value.

1.2.10.5 Haplotype

A haplotype is a set of alleles from closely linked loci inherited as a unit from the same parent. Traditional haplotyping simply specifies the parent from whom each child's allele is descended. However a more complete form of haplotyping can be undertaken that specifies the parental allele from which each child's allele is descended i.e. specifies grandparental source information (ibd).

Haplotyping was initially designed to make genetic data used in linkage analysis more informative. A locus is informative if parental allele can be inferred. For instance, a locus will be fully informative where both parents are heterozygotes and have no common alleles. However, for an individually uninformative loci, a highly polymorphic 'mega-locus' or haplotype can be constructed from a number of uninformative but closely linked loci. This combined 'mega-locus' is often informative at nearly all matings and this

Family pedigrees needed for linkage and association studies

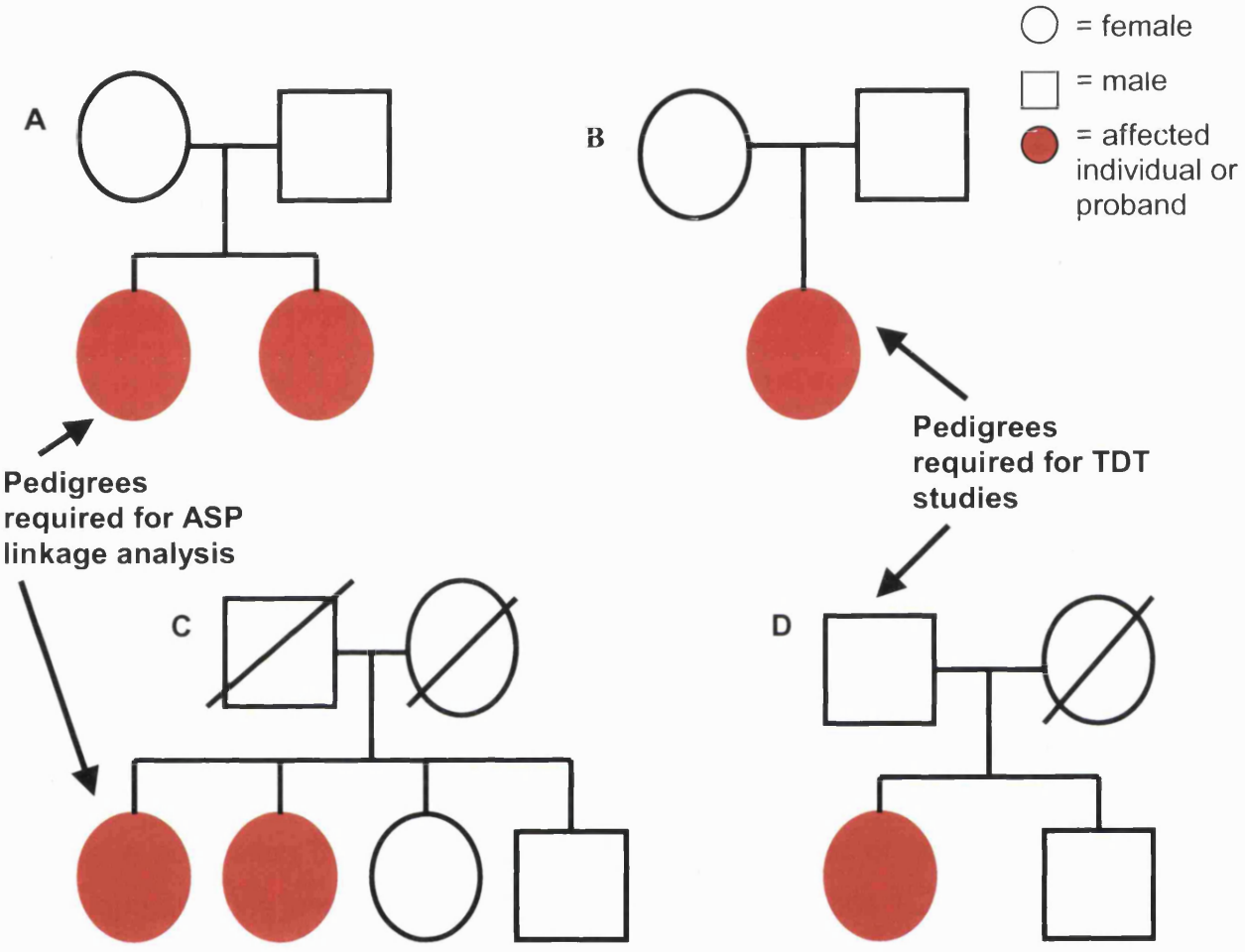


Fig 1.2.13 illustrates the family pedigrees required for affected sibling pair (ASP) linkage analysis and family based association studies or TDT analysis. Circles represent females, squares males and solid symbols in red represent affected individuals. Symbols with a diagonal line represent deceased individuals.
Pedigrees **A** and **C** represent the families required for linkage analysis and pedigrees **B** and **D** represent the families required for association analysis. Pedigree **A** is the ideal family to be collected for linkage studies because DNA availability from both parents improves the power of the study. The next ideal family for linkage analysis would be a one parent family. If neither parent remains alive then pedigree **C** allows the parental genotypes to be inferred from the offspring's genotypes. The least useful pedigree is that where DNA can only be collected from two affected siblings.
Pedigree **B** is the ideal family to be collected for TDT analysis.

principle will be used in third generation WGS using SNPs and DNA microarray technology.

Haplotyping can also be used to identify genotyping or data entry errors. Errors indicating in non-Mendelian inheritance are easy to detect but mis-typing a true 2/2 child as a 1/2 when both parents are 1/2 is difficult to detect. However, haplotyping across this locus should highlight the possibility that the child's typing was in error if for instance the haplotyping reveals a double recombination (one recombination on either side of the questionable allele).

1.2.10.6 Case-control analysis

It can be difficult to obtain an ideal data set for analysis especially in a clinically and genetically heterogeneous disease such as RA. It is tempting to devote a disproportionate amount of time to recruiting cases rather than controls. Often 'control' populations in genetic studies are derived from blood donors about whom there may be relatively little information. It is possible that the controls and cases may not come from the same population leading to potential sources of error. Alternatively, the individuals may come from the same population but this population may be composed of non inter-mating subpopulations leading to population stratification (sections 1.2.10 and 3.3.2). Unless family trees are available or very detailed questions regarding grandparents are asked it can be difficult to be certain when recruiting an individuals for a study whether their genetic background is similar to that of the other cases or controls. For instance, recruitment of individuals classified as Caucasoid gives a scope for a wide range of genetic backgrounds, from Celts to Anglo-Saxons to those from the Southern Mediterranean. Detailing genetic background can be time consuming and such questions may lead to fewer numbers recruited. However, case control studies are efficient with regard to the ratio of cases to controls genotyped. Recruitment is easier than for ASP analysis or TDT analysis. Provided the control populations used are

An example of a LOD score curve

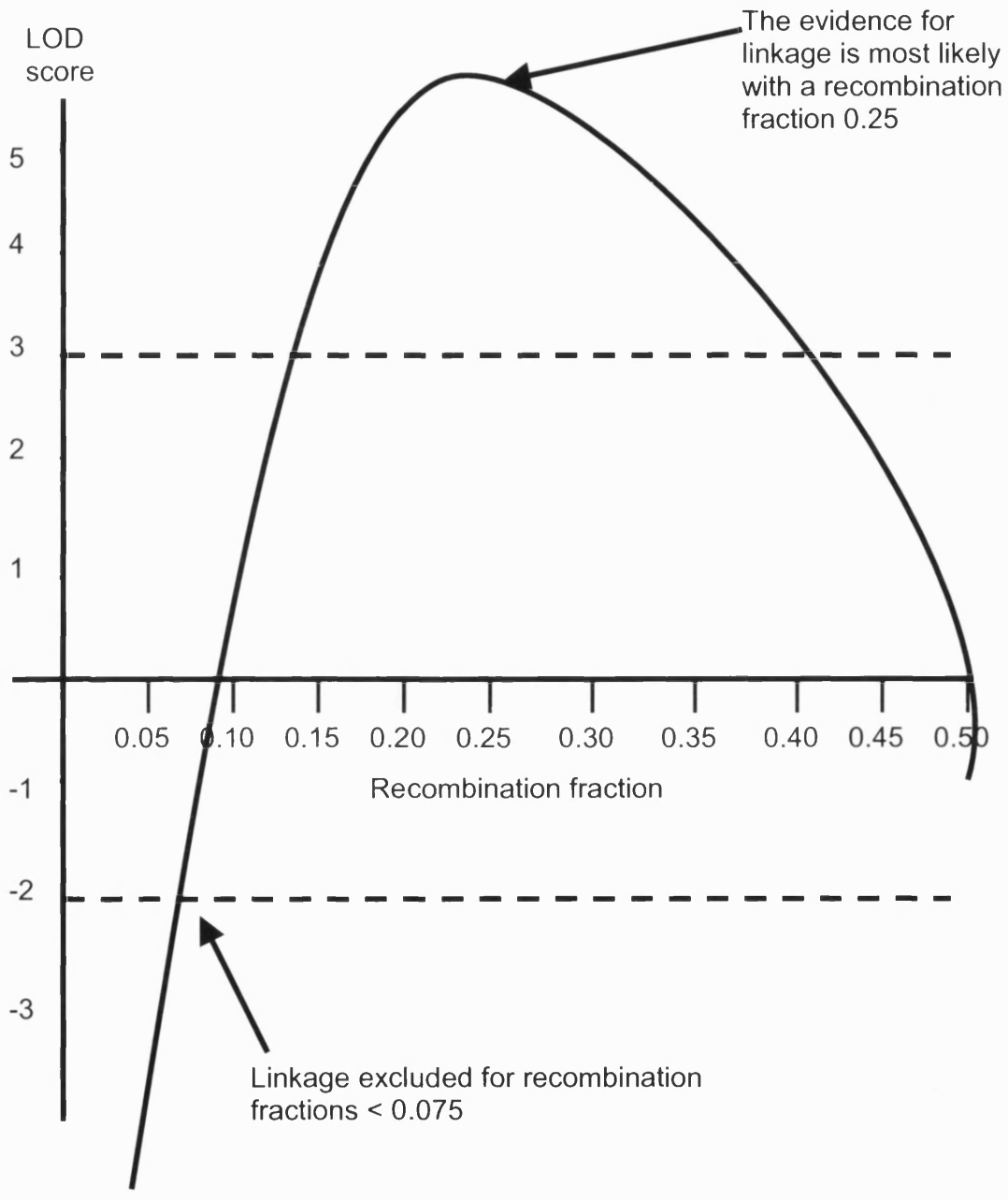


Fig 1.2.11 illustrates the relationship between a LOD score and recombination fraction in linkage analysis. When the LOD score at a marker is below –2, linkage at the marker is effectively excluded, if the LOD score at a marker is above 3.6 significant linkage is likely. Between a LOD of 2.2 and 3.6 linkage is suggested and between a LOD of 0.8 and 2.2 there is nominal evidence for linkage.

appropriate case control studies can be a useful additional method of investigation. Once families have been used to undertake haplotype analysis these haplotypes can be investigated in case control studies.

Chapter 1 – section 3

Chapter1, section 3

Laboratory procedures and techniques

The importance of statistics in human genetic research is clear and there are strong historical links between genetics and statistics. However, many of the advances in the science of genetics in recent decades are largely due to the emergence of new laboratory techniques. These techniques have transformed the approach to fundamental and applied molecular biology. The ability to amplify specific segments of DNA made possible by the polymerase chain reaction (PCR) is a corner stone technique that should be included in this category. PCR is an integral part of most genetic studies and was used as the basis for all the experiments described in this thesis.

1.3.2 Polymerase chain reaction (PCR)

PCR was invented by Kary Mullis in the mid-1980s (Mullis *et al.*, 1986) and was originally applied to the amplification of human β-globin DNA for the prenatal diagnosis of sickle-cell anaemia (Saiki *et al.*, 1988; Saiki *et al.*, 1985). Specific DNA sequences can be synthesised in vitro using two oligonucleotide primers that hybridize to opposite DNA strands flanking the region of target DNA. A repetitive series of cycles involving denaturation, primer annealing, and the extension of the annealed primers by DNA polymerase results in the exponential accumulation of a specific fragment the size of which is defined by the 5' ends of the primers (fig 1.3.1). Primer extension products synthesised in one cycle serve as a template in the next, therefore the number of target-DNA copies approximately doubles with each cycle. Twenty cycles of PCR yield about 2^{20} amplification. The purification of highly thermostable DNA polymerase from *Thermus aquaticus* (*Taq* DNA

polymerase) and the introduction of the thermal cycler has enabled the automation of PCR. In the original experiments the PCR process involved moving the samples between heat sources of differing temperatures and the addition of further DNA polymerase each cycle because of the thermal lability of the *E. coli* DNA polymerase used at that time.

DNA polymerase enables the addition of a dNTP at the free 3' hydroxyl group of the synthetic strand. This strand is complementary to the homologous strand. The annealing temperature is critical for ensuring that there is complementary binding of the oligonucleotide primer to its complementary sequence. Any mismatch at the 3' end of the primer will mean that the 3' hydroxyl group will not be recognised by the DNA polymerase as a site for new DNA synthesis.

Many different factors and techniques can alter the outcome of a PCR reaction. Varying the reaction parameters (e.g. changing the Mg^{2+} concentration or temperature cycling profile) can alter the specificity and yield of the reaction. For any given pair of oligonucleotide primers an optimal set of conditions can be established but no single set of conditions will be optimal for all possible reactions. Higher temperatures and Mg^{2+} concentrations increase the overall stringency of the reaction minimising the extension of primers that are mismatched with template. The concentration of the target sequence in the genomic template also influences the homogeneity of the PCR product (Saiki *et al.*, 1988). Non-target fragments can be produced if there is little target sequence. If there is insufficient *Taq* enzyme in the later cycles annealed primer-template complexes may not be fully extended in a single cycle period and a plateau effect can occur resulting in reduced efficiency. Other factors such as re-association of the template strands when the product concentration is high may also contribute to this effect. The error rate for *Taq* polymerase PCR is in the region of 1 in 10^4 cycles which does

not pose a problem for most applications apart from sequencing (Goodenow *et al.*, 1989; Tindall and Kunkel, 1988).

The possibility of contamination in the amplification reaction has broad implications for both research and diagnostic applications. Given the capacity of PCR to synthesise millions of DNA copies, contamination of the sample reaction with either the product of a previous reaction (product carryover) or the material from an exogenous source is a potential problem. Careful laboratory procedure include the pre-aliquoting of primers, the use of specific pipettes for loading PCR products into gels and the physical separation of the reaction preparation from post-PCR product. Minimising the number of reaction cycles also can reduce the chance that a rare contaminating template DNA will be amplified. 'Blank' lanes where no DNA template is added to the reaction mix must also be included to detect potential contamination.

Primer selection is very important as primers more than anything else will determine the success or failure of an amplification reaction. Certain guidelines can help with their design.

- Generally primer length should be between 18 –24 bases in length
- Where possible select primers with random base distribution and avoid primers with repetitious stretches of CCCC or AAAA etc. particularly at the 3' end.
- Check primers against each other for complementarity. Especially avoid primers with 3' overlaps to reduce the incidence of 'primer dimer'. 'Primer dimer' is an amplification artefact and occurs if the primers are complementary to each other or if many cycles of amplification are performed on a sample containing only a few initial copies of DNA template. The resulting concatenation is an extremely efficient PCR template and if amplified at an early cycle can overwhelm the reaction.

- Avoid sequences with significant secondary structure, particularly at the 3' end.
- Try to keep the GC content of the primer at a fixed level comparative to the AT content. The nucleotides cytosine and guanine are bound together by triple hydrogen bonds whereas thiamine and adenine are bound by double hydrogen bonds. Triple hydrogen bonds require more energy to break them than double hydrogen bonds. The melting temperature of the ds-DNA will therefore vary accordingly depending upon the GC versus AT content.

The specificity and yield of an amplification depend largely upon the concentration of Mg^{2+} and annealing temperature. Concentrations of 1.5 – 2 mM are optimal with 200μM of each dNTP. Optimal temperatures depend upon the *Taq* used and while some *Taq* enzymes are designed to work immediately others require a 'hot start' e.g. 10 minutes at 94°C. Avoiding contamination, optimising the conditions and using the correct *Taq* enzyme are of paramount importance.

PCR is one of the integral steps in many basic molecular biological techniques. Sequencing of the whole genome (Consortium, 2001) would not have been possible without PCR. Other applications include: the amplification of microsatellites (section 1.2.7.1), the identification of SNPs by direct sequencing, the detection and quantification of gene expression by synthesising a cDNA PCR template from an mRNA transcript using a reverse transcriptase and the ability to engineer or modify DNA fragments (e.g. producing primers with fluorescent tags attached). The process has been increasingly steam-lined in recent years. With the advent of DNA microarray technology high throughput will be possible and smaller quantities of DNA will be required.

1.3.3 Sample preparation

Sample preparation is also of great importance. Whatever the sample, the aim is to obtain a high DNA yield as free from contamination as possible and high molecular weight DNA rather than degraded DNA. DNA can be extracted from whole blood (using a variety of methods –appendix 5.3), from clinical swabs, synovial fluid, plucked hairs, tissue etc. If contamination with cell debris occurs whilst DNA is being extracted from whole blood the yield will be poor and the subsequent experiments difficult. Contamination can lead to difficulty with PCR and later with the assignment genotypes etc. Once whole blood is taken from the patient it is either frozen directly or frozen after the blood has been spun down and the leukocytes have been separated from the red cells. If whole blood has been frozen for more than a few months it can be more difficult to extract DNA and the yield may be reduced. Spectroscopy is performed to determine the DNA yield and purity of the sample (appendix 5.3).

Once extracted, conservation of DNA by careful husbandry is the next important step. Large quantities of DNA are frequently required for studies especially when undertaking genome wide linkage analysis where multiple genotypes are needed for each individual (section 2.2.3.3). It is unreasonable to expect volunteers to give more than about 20mls of blood at one time and there are problems, both ethically and logistically, in approaching an individual on more than one occasion for blood. Therefore ways must be found to conserve DNA or produce more DNA from the original sample. As technology develops methods to conserve DNA have improved but it is still necessary to be able to generate additional stocks of DNA where patient collections have taken considerable time and resources to put together. The Arthritis Research Campaign (ARC) UK family collection in RA has been recruiting RA families for about eight years. Two types of families are collected and include ASP families where DNA is available from two or more siblings with RA and also their first-degree relatives. The most useful

family of this type has when DNA available from both parents. The second type of family are singleton families where only one individual has RA but DNA is available from both parents. The ASP families are used for linkage studies and the singleton families are valuable for candidate gene research. Both types of families take time to collect; many individuals need to be approached, examined and bled and these families are relatively uncommon (especially the ASP families).

The ARC UK RA family collection holds labelled DNA stock in -70°C freezers but also has immortalised cell lines so that additional family DNA can be generated. Generating this additional disease DNA is time consuming and expensive therefore the conservation of DNA stocks remains important. DNA can be amplified by PEP PCR. This is a PCR amplification reaction which randomly amplifies DNA segments using randomly generated PCR primers (appendix 5.3). The DNA produced by this method is less robust than DNA produced directly from peripheral blood extraction and storage at temperatures other than -70°C for any length of time or frequent changes in temperature by freeze-thawing increases the likelihood of poor PCR results. The PCR amplification reaction is quite time consuming and the reagents expensive but it is a useful way of extending the capacity of a small quantity of DNA if multiple genotypes are required. It is very important when diluting DNA produced by this method that the sterile water used for the dilution is thoroughly mixed with the PEP DNA. If water is added and the resulting dilute DNA is then aliquoted into Costar plates for subsequent PCR reactions unless thorough mixing has taken place the DNA concentrations may vary between the different aliquots. This is especially noticeable between the first and last aliquots and can lead to PCR problems such as difficulty in assigning genotypes.

DNA is less robust to changes in temperature etc. when diluted. PCR reactions are often performed with dilutions of 10ng/μl therefore, repeated

freeze-thawing should be avoided and the dilute DNA should not be kept for any more than a few days at 4°C in a fridge. Equally, it should never be kept out on a laboratory bench for longer than necessary. This applies to all the reagents used to make up a PCR reaction mix. All should be kept on ice, especially the *Taq* polymerase.

Avoiding frequent freeze-thaw cycles for the dilute DNA or primers requires planning when undertaking a large study such as a genome wide linkage analysis. If many genotypes are required for many individuals, two approaches can be taken. Either a single primer can be used for all the DNA samples in one go (avoiding freeze-thawing the primer) or all the primers (~ 450) can be studied in one box of DNA at a time. A 'box' of DNA includes the 90 or so DNA samples that can be accommodated in each Costar plate. This second method can lead to primer degradation therefore the first approach is often used (especially as once a microsatellite has been amplified and separated for all individuals single point analysis of that primer could be undertaken).

The number of DNA freeze-thaw cycles can be reduced if the DNA is diluted to the necessary working concentration (often 10ng/μl) prior to aliquoting small volumes into Costar plates and then freezing. The volumes of the diluted frozen DNA need to be large enough to allow for 20 – 25 PCR reactions. This will be enough to amplify the 15 – 20 DNA fragments per primer set and allow for any PCR failures to be repeated. The number of primers per set will be determined by the number of microsatellites that can be run simultaneously on a polyacrylamide gel (which in turn will depend upon their size and fluorescent tag). For each set of microsatellites, the diluted DNA can be thawed and then aliquoted into suitable volumes for a single PCR reaction. The PCR mixes for each microsatellite can then be made, added to the DNA and a PCR reaction performed. Samples from each completed PCR reaction are then tested using an agarose gel. If any set falls

below a certain standard the PCR reaction is redone. The standards may vary depending upon circumstances but for the data presented in this thesis the standard used was 90 per cent PCR success (or more) on an agarose gel otherwise the PCR was repeated before the samples were run on a polyacrylamide gel. Once all microsatellites in a set have been successfully amplified the PCR products are diluted, mixed together and then separated by electrophoresis on 6% polyacrylamide gels using an ABI 373 semi-automated DNA sequencer (Applied Biosystems, Warrington, UK) over three hours or on 4% polyacrylamide gels and an ABI 377 DNA sequencer over two hours (section 2.3.3.7, figs 2.2.1 and 2.3.2). The optimal dilutions are determined by running a fluorescent test gel of diluted PCR products. Once the amplified PCR products have been successfully separated they are sized using the program GENESCAN™672 (version 2.1) (Applied Biosystems, Warrington, UK) and genotypes semi automatically assigned using the program GENOTYPER™ (version 1.1.1) (Applied Biosystems, Warrington, UK) (figs 2.2.3 and 2.3.3). A PCR product from a DNA reference sample (CEPH 1347-02) is included on every gel to monitor possible gel-to-gel variation. This sequence is applied to each set of primers.

Chapter 2 – section 1

CHAPTER 2: LINKAGE ANALYSIS

Systematic linkage screening of the entire genome and a replication study

- **Introduction**

Linkage analysis is an integral part of the search for susceptibility genes. Only the major histocompatibilty complex (MHC) on chromosome 6 has been consistently linked to and associated with RA susceptibility (Ollier and Thomson, 1992)(section 1.2.3) although family studies suggest that this accounts for just one-third of the genetic susceptibility (Deighton *et al.*, 1989; Wordsworth, 1991). It is likely that a number of other susceptibility loci exist, each contributing less than the MHC to the total genetic component. Given the small increase in disease risk for siblings (sibling recurrence risk or λ_s = 6 –14) (Deighton and Walker, 1991) any non-MHC susceptibility genes are likely to be of relatively small effect. Systematic linkage screening of the entire genome has the potential to detect all disease susceptibility loci if appropriately powered. However, it is necessary to study very large numbers of affected sibling pair (ASP) families if adequate power to detect linkage to non-MHC genes is to be generated (Brown and Wordsworth, 1998). Recruitment is difficult in a late onset disease such as RA and may require international collaboration. Undertaking whole genome analysis on such large numbers of individuals is both expensive and time consuming. Therefore the strategy has been to perform genome-wide screens on manageable numbers of families aiming to study any overlapping regions of nominal linkage ($p<0.05$) reported in two or more independent data sets in

greater detail with larger numbers of independent ASPs (Holmans and Craddock, 1997).

Three genome-wide linkage screens studying ASP families with rheumatoid arthritis have been reported. The studies have varied in size and have included 41 Japanese (Shiozawa *et al.*, 1998), 97 European (Cornelis *et al.*, 1998) and 251 North American ASP families (Jawaheer *et al.*, 2001). Linkage to the MHC was confirmed ($p< 2.5\times10^{-5}$) in the European (ECRAF) and North American (NARAC) studies (Cornelis *et al.*, 1998; Jawaheer *et al.*, 2001). All three studies identified a number of non-MHC regions with evidence of nominal linkage ($p<0.05$) and a few regions with suggestive linkage ($p<0.001$). It is inevitable that many apparent linkages will be false positives (Lander and Kruglyak, 1995), hence, all findings need to be confirmed in independent data sets. Candidate regions likely to contain true RA susceptibility loci are those where nominal linkage ($p<0.05$) has been reported in two or more of these studies. The eight regions on chromosomes 3q, 8p, 12q, 14q, 16p, 16q, 18q and Xp (Cornelis *et al.*, 1998; Jawaheer *et al.*, 2001; MacKay *et al.*, 2002) where this has occurred are summarised in table 2.1.1.

No linkages outside the HLA region have been confirmed ($p<2.5\times10^{-5}$) in RA and the very large numbers of ASPs required to detect linkage to genes of small effect may mean that some regions of true linkage may have been missed. Hence, further genome-wide linkage screens studying ASP families with RA are of value. Larger studies are less prone to type 1 or type 2 error and any overlapping regions identified in another independent cohort of ASPs with RA will increase the likelihood that one of the candidate regions listed in table 2.1.1 contains a true RA susceptibility locus.

Once regions of nominal linkage have been identified (particularly overlapping areas) replication studies can be undertaken. These can be efficient in terms

of time and resources so larger numbers of individuals can be genotyped (Holmans and Craddock, 1997). Also with fewer markers studied the adjustment required to correct for multiple testing will be smaller (section 1.2.8.4).

Two studies are reported below, both undertaken in collaboration with the ARC Epidemiology unit, at Manchester University. The first is a replication study undertaken in 1998/99 before the publication of the North American genome screen (Jawaheer *et al.*, 2001). It was designed to confirm in a larger independent cohort of 368 ASPs 25 regions of nominal linkage as reported by the European genome screen (Cornelis *et al.*, 1998) (table 2.2.1). The second is a genome-wide linkage analysis of 182 UK ASP families (252 ASPs). This WGS was initiated because it was important to identify additional sites of nominal linkage in an independent cohort of ASPs with RA. It was undertaken prior to the NARAC study and included more that twice the number of ASPs studied by ECRAF (thereby improving power). Hence, the aim was two-fold: to determine whether there were any overlapping regions of nominal linkage in the independent cohorts of RA ASPs and secondly to identify novel regions of nominal linkage (which may not have been identified in the ECRAF cohort of only 114 ASPs). If the UK ASPs had been used as a replication cohort only any novel regions would not have been identified and maybe important areas for further study missed. Although the costs of a WGS (both in terms of time and money) are high these could be justified as only one moderate sized WGS had been published at that time. As more data becomes available efforts should be aimed at replicating already identified sites of nominal linkage with high density markers.

Chapter 2 – section 2

2.2 Linkage analysis of putative non-HLA rheumatoid arthritis susceptibility loci in 368 UK affected sibling pair families.

2.2.1 Abstract

Objective: Using an independent cohort of 368 United Kingdom sibling pairs affected by rheumatoid arthritis (RA), this study was designed to test for linkage to candidate susceptibility loci identified in the first systematic linkage screen of the whole genome in RA

Methods: Using 59 informative microsatellite markers 368 affected sibling pairs (ASPs), from 280 families, were genotyped. These microsatellites mapped to regions identified in the first genome-wide linkage study in RA. Markers mapping in the region of estrogen synthase were also analysed based on previous reports of linkage to this gene. Automated genotyping was performed using fluorescent PCR primers and ABI DNA sequencers. Statistical analysis was undertaken using ANALYZE for single point analysis and MAPMAKER/SIBS (version 2) for multipoint analysis.

Results: Markers on chromosomes 12, 15, and 21 (d12s95, CYP 19, d21s1252) showed evidence of nominal linkage ($p<0.05$) by single point analysis. However, no evidence of nominal linkage was found following multi-point analysis.

Conclusion: Consistent evidence of linkage was not found for any of the regions previously reported as potential non-HLA RA susceptibility loci in a large independent cohort of UK RA ASPs.

2.2.2 Introduction

The first systematic linkage screen of the whole genome of individuals with RA was published in 1998 and was undertaken using 114 RA affected sibling pairs (ASPs) from 97 families (Cornelis *et al.*, 1998). These ASPs were recruited from a number of European countries including France, Italy, Spain, Holland, Belgium. Linkage to HLA was confirmed ($p<2.5\times10^{-5}$) and nominal evidence of linkage ($p<0.05$) was obtained for a further 14 regions (19 markers) based on multipoint analysis and for a additional eight regions based on single point analysis (Cornelis *et al.*, 1998)(table 2.2.1). Two of the regions (18q22-23 and 3q13) contained loci previously linked to insulin-dependent diabetes mellitis (IDDM-6 and IDDM-9). These were investigated with additional markers in a second cohort of 194 European ASPs but further support for linkage was only found for the IDDM-9 region on chromosome 3 when all individuals were analysed together (Cornelis *et al.*, 1998).

A second much smaller study of 41 Japanese families suggested linkage with regions on chromosome 3, 8 and X (Shiozawa *et al.*, 1998). Unusually, linkage to HLA was not confirmed in this study which may demonstrate ethnic differences in susceptibility genes or may be as a result of the smaller number of ASPs studied. Linkage to the marker CYP 19 on chromosome 15 has been demonstrated in two large samples (225 and 107) of UK RA ASPs (John *et al.*, 1999). CYP 19 is a polymorphic tetranucleotide marker in intron D of the estrogen synthase locus. The evidence for linkage was strongest in patients with an age at onset that was >50 years.

The results of linkage analysis of candidate loci in a large independent cohort of UK RA ASPs are reported below. The candidate loci were identified from the ECRAF study but also include Cyp 19 following the reports of linkage above.

European whole genome screen in RA: results of single and multipoint analysis

Marker	Distance from telomere cM	Single point p value	Multipoint p value
D1S228	32.4	0.0035	0.0065
D1S238	206.7	0.022	0.099
D2S377	228.2	0.024	0.013
D2S2354	235	0.0043	0.0054
D3S1267	141	0.032	0.039
D5S422	163.9	0.033	0.045
D6S292	138.2	0.036	0.034
D12S99	13.9	0.0077	0.083
D12S95	97.7	0.0067	0.043
D13S170	65.4	0.039	0.015
D13S1315	105.2	0.00035	0.0037
D14S285	50	0.049	0.33
D16S420	43.2	0.047	0.039
D16S401	45.5	0.008	0.028
D18S57	63.2	0.018	0.033
D18S474	71.3	0.038	0.012
D18S68	94.4	0.032	0.02
D18S61	102.8	0.0016	0.0098
D18S469	109	0.0055	0.02
D20S864	0	003	0.11
D21S270	41.3	0.038	0.11
D21S268	44.9	0.025	0.13
D22S264	0.9	0.019	0.0098
DXS1068	56.2	0.044	0.27
DXS998	183.8	0.0078	0.019

Table 2.2.1 summarises the results of the genome wide linkage analysis undertaken using the European consortium of RA families (ECRAF) [Cornelis, 1998]. This WGS examined 97 families, containing 114 ASPs. Significant linkage to the HLA region was observed, as expected. This table shows the additional 14 regions identified that were compatible with nominal linkage ($P<0.05$ without correction for multiple testing). Results for single and multipoint analysis are shown. Nominal linkage was identified at more than one marker in some of the regions.

2.2.3 Methods

RA Affected Sibling Pairs

Three hundred and sixty-eight UK Caucasian sibling pairs, affected by rheumatoid arthritis, were studied. The affected sibling pairs, from 280 families, were identified from the Arthritis and Rheumatism Campaign United Kingdom (ARC-UK) National Repository of family material (www.http://arc.man.ac.uk). Table 2.2.2 documents the pedigree structure of the 368 ASPs and table 2.2.3 the number of siblings per family. The affected sibling pairs and their families were recruited following a nation-wide publicity campaign in newspapers and local radio. Rheumatologists throughout the UK also identified ASPs who were willing to enrol in the study. Multi-centre ethics committee approval was obtained. All family members were examined according to a standard protocol and agreed to a detailed structured interview regarding joint symptoms. A trained metrologist performed a joint examination to detect swelling, tenderness and deformity. Individuals with erosive disease were identified from hand radiographs (all reviewed by a rheumatologist) and rheumatoid factor (RF) status (ascertained using a particle agglutination test). Subjects were classified as having rheumatoid arthritis if they satisfied the 1987 American College of Rheumatology (ACR) criteria (Arnett *et al.*, 1988) modified for genetic studies (MacGregor, 1995) (section1.1.3 and table 1.1.3).

DNA from the affected individuals and their first-degree relatives was prepared from peripheral blood. Technicians employed by the ARC-UK National Repository performed all the DNA extraction according to standard techniques. Concentrated DNA, received from the ARC-UK National Repository on ice, was kept in labelled tubes at -70°C at the Wellcome Trust Centre from Human Genetics (WTCHG) in Oxford or the Arthritis Research Campaign Epidemiology Unit (ARC-EU) in Manchester. It was diluted with

A fluorescent gel from the replication study

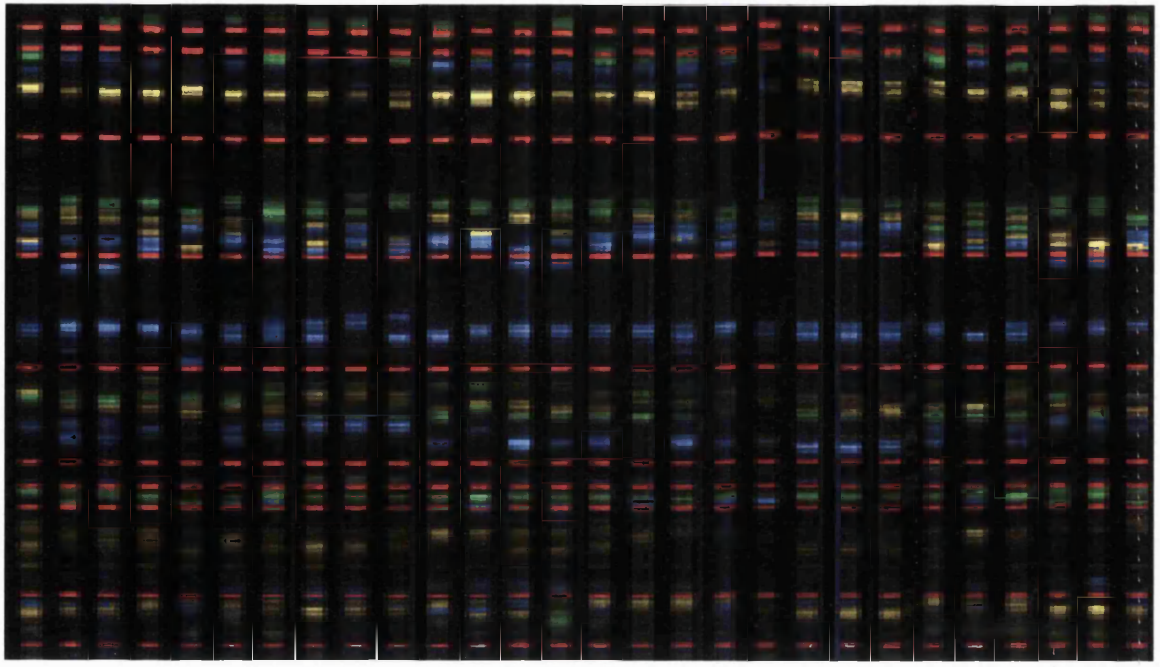


Fig 2.2.1 shows a fluorescent gel from an ABI 373 having been run through the software package Genescan™. The red lines are marker lines allowing the software package Genotyper™ to determine the size of each microsatellite allele relative to the known size of the markers. The yellow, blue and green lines are the fluorescently tagged microsatellites. Each column represents one individual and up to 20 microsatellite markers per person can be separated by PCR provided the fragments of a similar size are tagged by a different fluorescent colour.

sterile water to 10ng/μl before use. Dilute DNA was kept at 4°C for the duration of the study.

Microsatellite markers

Eighteen regions with nominal evidence of linkage ($p<0.05$) identified by single or multi-point analysis in the European RA WGS (Cornelis *et al.*, 1998) were included for study. A 16cM region containing the CYP 19 locus on chromosome 15 was also included. Fifty-nine suitable and informative microsatellites within these regions were identified (Table 2.2.4). Where possible the same microsatellite markers as used in the European RA WGS were chosen but when this was not feasible markers nearby were included. The microsatellites chosen needed to complement each other with regards to their PCR product size so that they could be run together on polyacrilamide gels (12 –15 microsatellites per lane).

Microsatellite genotyping

Semi-automated analysis of microsatellite genotypes was carried out in two centres; the Wellcome Trust Centre for Human Genetics (WTCHG), Oxford and the Arthritis Research Campaign Epidemiology Unit (ARC-ERU), Manchester. The microsatellites were amplified by PCR using fluorescently labelled primers. Reactions were carried out in 96 well plates (Costar) in 10μL reactions consisting of 50ng DNA, 400nM each primer, 50μM each dNTP, 1-3mM $MgCl_2$ and 0.2 units DNA polymerase (Bioline, UK) in the manufacturers NH_4 buffer, overlaid with liquid paraffin. The cycling conditions were 32 cycles of denaturation (1minute, 95°C), primer annealing (30 seconds, 55°C - 60°C) and extension (30 seconds, 72°C). Annealing temperatures and $MgCl_2$ buffer concentration were optimised for each primer. Amplification of each microsatellite was performed separately and the PCR products were then combined into pools of 10 – 12 markers before loading. Products were separated by electrophoresis either on 6% polyacrylamide gels using ABI 373 semi-automated DNA sequencers (Applied Biosystems,

Genotyper for the marker D21S1252

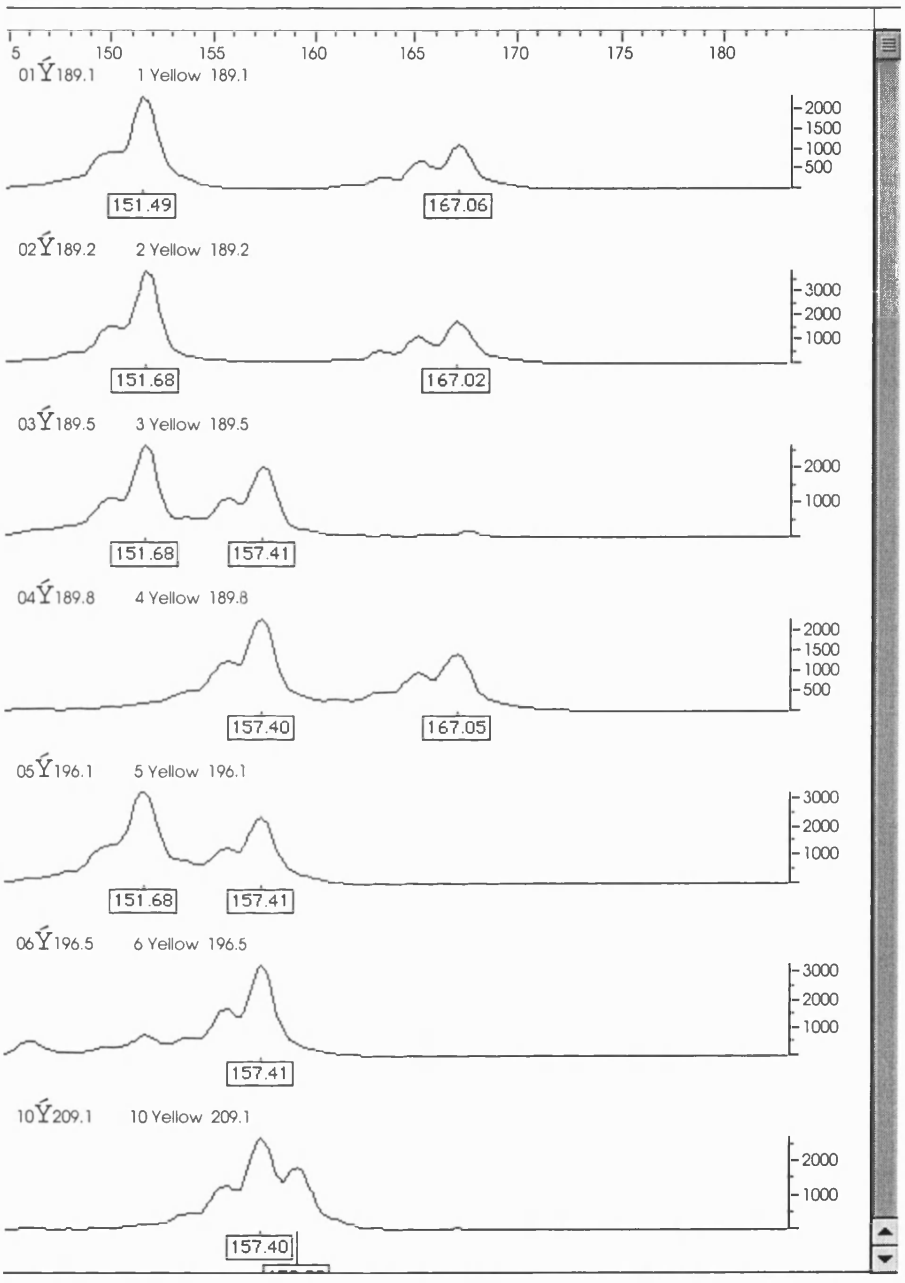


Fig 2.2.2 Shows the Genotyper™ output for the marker D21S1252 using the families 189 and 196. Once the microsatellite D21S1252 has been amplified by PCR and the fragments separated by gel electrophoresis using the ABI 373, the output is run through the software package Genescan™ to appropriately align the PCR fragments with the red marker fragments and then the result is run through Genotyper™. This sequence allows for the size of each PCR fragment (or microsatellite allele) to be identified.

Warrington, UK) over three hours (WTCHG) or on 4% polyacrylamide gels and ABI 377 DNA sequencers over two hours (ARC-EU). Test gels were run to ensure the PCR dilutions were optimised prior to separating PCR products by electrophoresis. Products were sized using the program GENESCAN™672 (version 2.1) (Applied Biosystems, Warrington, UK) and genotypes semi automatically assigned using the program GENOTYPER™ (version 1.1.1) (Applied Biosystems, Warrington, UK) (figs 2.2.1 and 2.2.2). A PCR product from a DNA reference sample (CEPH 1347-02) was included on every gel to monitor possible gel-to-gel variation.

Statistical analysis

Error checking:

To minimise data errors error-checking procedures were employed. Allele assignment by GENOTYPER™ was checked manually for all genotypes and the CEPH (Centre d'etude des polymorphisms humaines) control sample was used to ensure the consistency of allele assignment. The program GAS (Version 2) (A. Young, unpublished) was used to convert the size data into discrete allele numbers and samples not following Mendelian patterns of inheritance were identified in PEDCHECK (O'Connell and Weeks, 1998) and removed from the data set.

Non-parametric analysis:

Allele frequencies were calculated from all of the scored genotypes using the software packages DOWNFREQ (J. Terwilliger, unpublished) (WTCHG) or SPLINK (version 1.05, David Clayton, MRC Biostatistics Unit, Cambridge) (ARC-EU). Sibling pair analysis methods were then used to test for excess allele sharing between affected siblings. Non-parametric single point and multipoint analysis was implemented in MAPMAKER/SIBS (version 2) (Kruglyak and Lander, 1995). Multiple sibships were given a conservative weighting of 2/n to account for their lack of independence (Meunier *et al.*, 1997).

DNA availability	Number of families
Both parents	33
One parent	42
Siblings only:	205
Between 3 and 8 siblings per family	113
2 siblings (only)	92

Table 2.2.2 shows the pedigree structure of the families included in the replication study. For every affected sibling pair (ASP) additional first-degree relatives were recruited in order to ascertain the identity by descent of each allele studied. This table documents the availability of DNA from parents or additional siblings for each ASP family used in the UK replication study. If parental DNA was unavailable, unaffected siblings were recruited and genotyped if possible allowing the missing parental genotypes to be inferred. In total 958 individuals from 280 families were genotyped.

Affected sibling pairs per family	Number of families
2 ASPs	254
3 ASPs	20
4 ASPs	5
6 ASPs	1

Table 2.2.3 shows the number of ASP families recruited for the replication study and the number of ASPs per family. Some of the families included were made up of multiple sib-ships of between 3 and 8 RA affected siblings and the number of multiple sib-ship families are documented here.

Analysis of markers on the X chromosome was carried out using MAPMAKER/SIBS modified for X-linked loci (Cordell *et al.*, 1995). The program calculates sharing between sister-sister, brother-brother and sister-brother affected sib pairs.

Both multipoint and single point analyses have been presented. The suggested thresholds of LOD ≥ 3.6 ($p \leq 2 \times 10^{-5}$) for significant linkage and LOD ≥ 2.2 ($p \leq 7.4 \times 10^{-4}$) for suggestive linkage have been used (Lander and Kruglyak, 1995). Nominal evidence of linkage was taken to be LOD ≥ 0.8 ($p < 0.05$) by either multipoint or single point analysis (section 1.2.8.4).

2.2.4 Results

The results of the non-parametric, single point and multipoint analysis implemented in MAPMAKER/SIBS are shown in table 2.2.4. Fifty-nine informative microsatellite markers were genotyped successfully in more than 85% of individuals. Nominal evidence for linkage was identified for three markers by single point analysis: D12S95 ($p < 0.05$), CYP19 ($p < 0.05$) and D21S1252 ($p < 0.02$). Multi-point analysis, using MAPMAKER/SIBS, did not provide further evidence for linkage.

2.2.5 Discussion

Three markers (D12S95, CYP19 and D21S1252) on chromosomes 12, 15 and 21 showed nominal evidence for linkage ($p < 0.05$) by single point analysis. Multi-point analysis did not provide further evidence for linkage. The marker D12S95 (chromosome 12) was also included in the European WGS and showed nominal evidence of linkage in that study by both single

Replication study: results table

Marker	Distance from T (cM)	LOD (SP)	p value (SP)	LOD (MP)	Marker	Distance from T (cM)	LOD (SP)	p value (SP)	LOD (MP)
D1S2667	26.9	0	0.5	0	D13S 1315	105.2	0.011	0.41	0
D1S 2740	32.2	0.08	0.27	0	D13S 285	112.8	0	0.5	0
D1S 434	32.2	0	0.5	0	D14S 276	47	0.1	0.25	ND
D1S228	32.4	0	0.5	0	D15S 994	43	0	0.5	0
D1S 507	36.2	0.43	0.08	0	D15S 119	51	0	0.5	0
D1S2672	38	0	0.5	0	CYP19	51	1.2	0.02	0
D1S 196	186.4	0.021	0.38	0	D15S 992	51	0.14	0.27	0
D1S 238	206.8	0.014	0.4	0	D15S 962	55	0.18	0.2	0
D1S 413	216.5	0.18	0.18	0	D15S 117	59	0	0.5	0
D2S 2308	234.5	0.006	0.43	0	D16S 420	43.2	0.057	0.3	0
D2S 2354	235	0.023	0.37	0	D16S 401	45.5	0	0.5	0
D3S 3649	136.7	0	0.5	0	D16S 3100	50.8	0.09	0.26	0
D3S 3567	139.9	0	0.5	0	D16S 3040	110	0.017	0.39	0
D3S 1267	141	0	0.5	0	D18S 1102	63.1	0.13	0.22	0
D3S 3646	141.1	0	0.5	0	D18S 465	98.9	0	0.5	0
D3S 1589	143.8	0.27	0.2	0	D18S 488	104.6	0.25	0.14	0
D5S 410	156	0	0	ND	D18S 469	109	0	0.5	0
D6S 1684	144	0	0.5	ND	D18S 462	118	0	0.5	0
D6S 144	155.3	0	0.14	ND	D20S 117	2.9	0.16	0.19	ND
D12S 99	13.9	0	0.5	0	D21S 1256	8.6	0.52	0.06	0
D12S 85	65.6	0.04	0.33	0	D21S 1914	23	0.014	0.4	0
D12S 368	67.3	0.36	0.1	0	D21S 263	31.4	0	0.5	0
D12S 83	76.5	0.11	0.24	0	D21S 1254	34	0	0.5	0
D12S 326	87.6	0.08	0.27	0	D21S 1252	38.7	0.82	0.02	0
D12S 351	97.1	0.002	0.47	0	D21S270	41.3	0.2	0.2	0
D12S 95	97.7	0.59	0.05	0	D21S 268	44.9	marker	failed	
D12S 327	99.6	0	0.5	0	D21S 266	49.9	0.04	0.33	0
D12S 346	106	0	0.5	0	D22S 420	0	0	0.5	ND
D13S 170	65.4	marker	failed		DXS 1068	56.2	0	0.56	ND
D13S 1265	101.7	0	0.5	0	DXS1200	179.8	0	0.5	ND
					DXS 998	183.8	0	0.5	ND

Table 2.2.4 lists the results of the replication study including the markers used, LOD score, and p value for single point and multipoint analysis (where appropriate). ND = not done. Single point analysis on 368 UK-ASPs was performed for each marker implemented in ANALYZE and multipoint for each locus implemented in MAPMAKER SIB (LOD mp).

point analysis (p=0.0067) and multi-point analysis (p=0.043). D21S1252 on chromosome 21 maps closely to the markers D21S270 (2.6cM telomeric) and D21S268 (6.2cM centromeric) used in the European WGS. These European WGS markers showed evidence for nominal linkage by single point analysis ($p<0.05$) but not multi-point analysis. CYP19 on chromosome 15 has previously shown nominal evidence of linkage in two stage testing, although there was some overlap in the families used in the two investigations (John *et al.*, 1999).

It is not surprising that only three of the 14 regions identified by the European genome screen were reproduced in this replication study. Linkage analyses undertaken in other complex diseases reveal the same lack of consistent results when studying independent family collections. Examples include studies of insulin dependent diabetes mellitus (IDDM) (Concannon *et al.*, 1998; Davies *et al.*, 1994; Luo *et al.*, 1996) and multiple sclerosis (Ebers *et al.*, 1996; Haines *et al.*, 1996). This inconsistency has been attributed to a number of factors including clinical and genetic heterogeneity, the use of different markers and analytical methods and poorly powered study designs unable to detect weak genetic effects. It is also accepted that a large proportion of apparent linkages identified by genome-wide linkage analysis will be false positives requiring confirmation of results in independent data sets (Lander and Kruglyak, 1995). As the most efficient strategy is thought to include an initial WGS screen performed with relatively few ASPs and wide marker spacing, followed by the more detailed study of all loci with nominal evidence for linkage ($p<0.05$) in a larger data-set (Holmans and Craddock, 1997), inconsistent results are inevitable. Candidate regions likely to contain true RA susceptibility loci are those where nominal linkage ($p<0.05$) has been reported in two or more of these studies and where both single and mulit-point analysis show evidence for linkage. The single point linkage to the three regions identified on chromosomes 12, 15 and 21 has not been supported by multi-point results but the power to detect weak linkage effects

is limited in a study of this size. Therefore these regions cannot be formally excluded from containing relevant genes and warrant further investigation.

Clinical heterogeneity in RA is well recognised and differences in phenotype between study populations may influence results. The two studies recruited ASPs in slightly different ways and from different regions within Europe. This may have lead to phenotypic differences between the disease cohorts and some variation in the results. All individuals included in the ARC-UK National Repository of family material were recruited following a detailed structured interview and joint examination undertaken by a trained metrologist. Hand radiographs were reviewed looking for erosive disease and rheumatoid factor status was determined. The ASPs recruited by the European WGS fulfilled 1987 ACR criteria for RA as reported by one of the rheumatologists participating in the study or the clinician in charge of the patient who completed a standardised questionnaire (Cornelis *et al.*, 1998). However some family members would not have been seen or examined, and is therefore possible individuals with RA may have been missed within the family group.

A variety of statistical approaches are available to analyse data generated by linkage studied of this kind. The European WGS used the programme ANALYZE (J Terwilliger, unpublished) for single point analysis and GENEHUNTER 1.1 (Kruglyak *et al.*, 1996) for multi-point linkage analysis. In addition, for the few families with multiple affected sib-ships 'all possible pairs' were used as if they were independent ASPs. The UK replication study used the program MAPMAKER/SIBS (version 2) (Kruglyak and Lander, 1995) for single and multi-point analysis and multiple sibships were given a conservative weighting of 2/n to account for their lack of independence (Meunier *et al.*, 1997). The use of different statistical programmes (section 1.2.9) and different approaches to the analysis of multiple sibships (section 1.2.7.6) may have lead to small differences in the results of either study.

MAPMAKER/SIBS (version 2) (Kruglyak and Lander, 1995) was designed for the analysis of small nuclear families or sibships and the method is based on the Lander-Green algorithm (Lander *et al.*, 1987) (section 1.2.9.1). Power calculations often overestimate the power of a study because the calculations typically assume that all families are fully informative. It is assumed that DNA is available from all parents, marker heterozygosity is 100% and all affected sibling pairs are independent (which is not the case in families with more than two affected siblings). Recruitment of complete affected sibling pair families is difficult in a late onset disease such as rheumatoid arthritis. Thus, parental DNA was available in 27% of the families studied. DNA from unaffected siblings was available in 55% of the remaining families allowing parental haplotypes to be inferred when not directly available.

Two other regions identified by the European WGS, 18q22-23 and 3q13 contain the IDDM-6 and IDDM-9 loci and were studied in a further 194 European ASPs from 164 families (Cornelis *et al.*, 1998). The region 18q22-23 did not reach significance in the larger cohort. Suggestive linkage to the chromosome 3 (IDDM-9) region was found in the combined data set but was only evident in HLA concordant RA pairs (p=0.001). Linkage was not significant in the remaining RA pairs (p=0.08). Candidate genes within the 3q13 region include genes coding for CD80 and CD86 which interact with T cell molecules CTLA4 and CD28 in the co-stimulatory pathway (Cornelis *et al.*, 1998). Although, no evidence for linkage in either region was identified by the UK replication study the region on chromosome 3 at 3q13 warrants further investigation as this area has been identified in two independent datasets. The region 18q22-23 is less likely to harbour important disease causing genes as significance was not reached in either the second independent European dataset (Cornelis *et al.*, 1998) or the UK replication study.

A second WGS has been reported, which was performed in a small number of Japanese families (n=41) (Shiozawa *et al.*, 1998). Multi-point linkage analysis revealed two regions with maximum LOD score >3 (significant linkage) on chromosome 1 and the X chromosome. The markers on chromosome 1 were D1S253 and D1S214 and mapped 16cM from D1S228 a marker with evidence of nominal linkage in the European study. The marker on the X chromosome was DX1232 mapping approximately 25cM from DSX998 another a marker with evidence of nominal linkage in the European study. These findings were not replicated in the UK dataset. Interestingly, significant linkage to the HLA-DR region was not detected in the Japanese families. This could be due to ethnic differences in HLA associations or more likely the relatively low power of such a small study to detect linkage. Both disease severity and HLA associations vary amongst different populations (section 1.2.3). All the UK families were Caucasoid and the linkage observed to the HLA-DRB1 region was similar to that seen in the European WGS (MacKay *et al.*, 2000).

RA is a chronic multifactorial disease and is thought to be genetically heterogeneous. A number of earlier studies of complex diseases have found the strongest evidence for linkage when confining the analysis to specific subsets of sibling pairs (e.g. HLA discordant versus HLA concordant), or by looking at interactions between loci. However, it is recommended that subset analysis should only be undertaken when there is some evidence for linkage in the whole data set prior to sub-analysis (Leal and Ott, 2000). Therefore, further analysis of the UK replication study was not performed in view of the lack of strong evidence of linkage overall.

In conclusion, the contribution of individual disease susceptibility loci is likely to be small in RA as in other complex diseases. Additional genome-wide linkage studies are required to replicate any evidence of linkage at specific loci already identified and generate additional regions where linkage may be

found. Hence, a WGS using the UK RA ASP families should be undertaken. Many of the loci identified by these studies will be false positives and large numbers of ASPs will ultimately be required to replicate results. World-wide collaboration may be required if suitable families are to be recruited in sufficient numbers.

Chapter 2 – section 3

2.3 Whole genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 United Kingdom affected sibling pairs.

2.3.1 Abstract

Objective: To undertake a systematic whole genome screen to identify regions exhibiting genetic linkage to rheumatoid arthritis (RA).

Methods: Two hundred and fifty-two RA affected sibling pairs, from 182 United Kingdom families, were genotyped using 365 highly informative microsatellite markers. Microsatellite genotyping was performed using fluorescent PCR primers and semi-automated DNA sequencing technology. Linkage analysis was undertaken using MAPMAKER/SIBS for single point and multipoint analysis.

Results: Significant linkage was identified around the MHC region on chromosome 6 (maximum LOD score = 4.7 (p=0.000003) at marker D6S276 1cM from HLA-DRB1). Suggestive linkage ($p < 7.4x10^{-4}$) was identified on chromosome 6q by single and multipoint analysis. Ten other sites of nominal linkage ($p<0.05$) were identified on chromosomes 3p, 4q, 7p, 2 regions of 10q, 2 regions of 14q, 16p, 21q and Xq by single point analysis and three sites (1q, 14q and 14q) by multi-point analysis.

Conclusion: Linkage to the MHC region was confirmed. Eleven non-HLA regions demonstrated evidence of suggestive or nominal linkage but none reached the genome-wide threshold for significant linkage ($p = 2.2 x10^{-5}$). Previous genome screens have tentatively suggested six of these regions to be involved in RA susceptibility.

2.3.2 Introduction

Genome-wide linkage analysis can generate additional regions where linkage may be found and can replicate any evidence of linkage at specific loci already identified. A third WGS in RA was published in 2001 using 251 North American ASP families collected by the North American RA consortium (NARAC). Linkage to the MHC was confirmed (i.e. $p<2.5\times10^{-5}$) (Jawaheer *et al.*, 2001) and non-MHC regions with evidence of suggestive linkage ($p<7.4 \times 10^{-4}$) or nominal linkage ($p<0.05$) by multipoint analysis were also identified. As discussed above many apparent linkages will be false positives and confirmation of all findings in further independent datasets is important (Lander and Kruglyak, 1995). Table 2.1.1 summarises the eight overlapping regions where nominal linkage has been reported in two or more data sets (Cornelis *et al.*, 1998; Jawaheer *et al.*, 2001; MacKay *et al.*, 2002). These are important candidate regions for including true RA susceptibility loci. However, some regions of true linkage are likely to have been missed because most whole genome screens have relatively low power to identify linkage to genes of small effect (Brown and Wordsworth, 1998).

A WGS was performed on a large well-characterised cohort of United Kingdom ASP families with RA was undertaken to identify novel susceptibility loci and confirm linkage to previously identified regions.

2.3.3 Methods

2.3.3.1 Recruitment

Two hundred and fifty-two UK Caucasian sibling pairs, affected by rheumatoid arthritis, were studied. The ASPs, from 182 families, were identified from the Arthritis and Rheumatism Campaign United Kingdom

Flow chart for genome wide linkage analysis

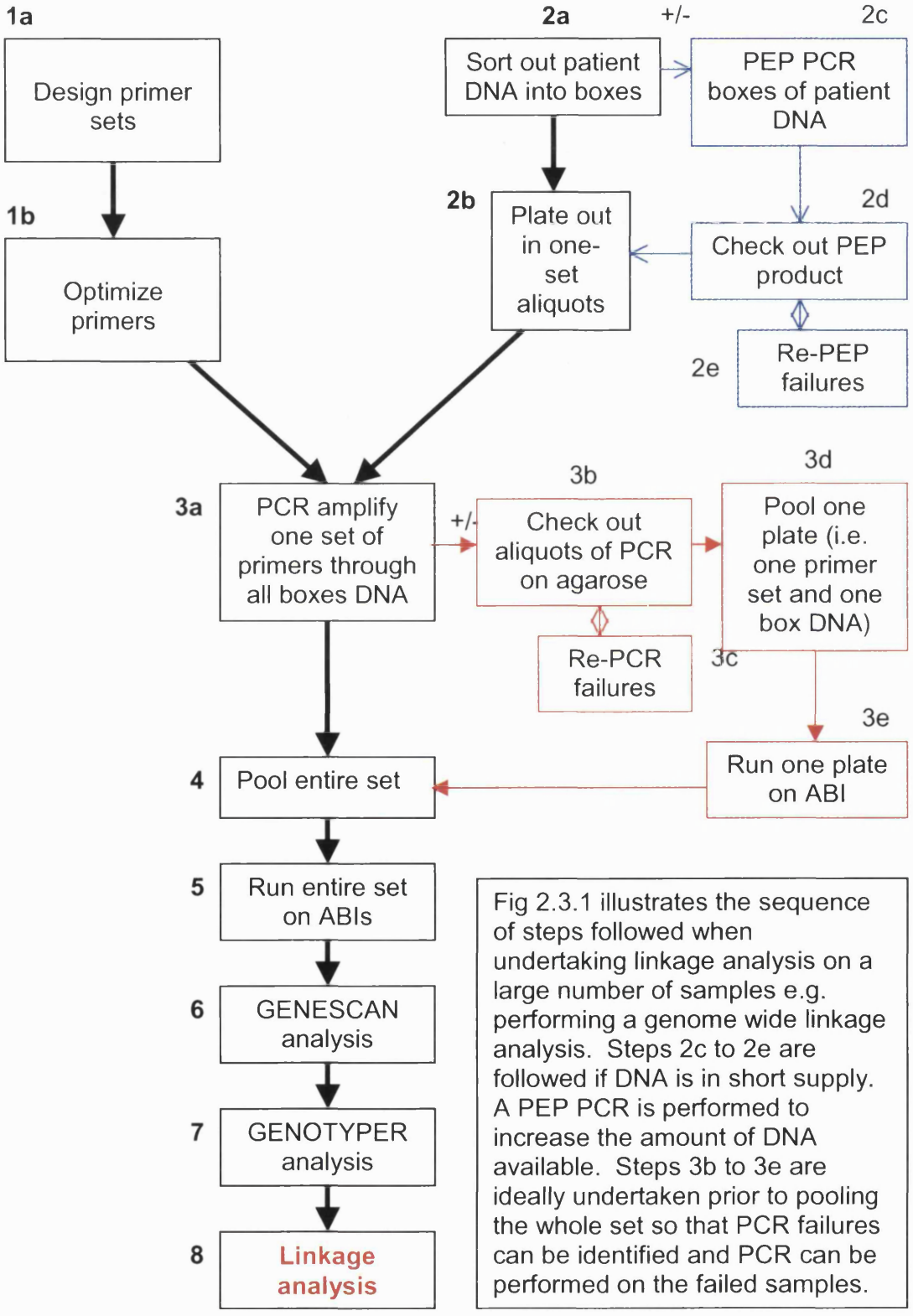


Fig 2.3.1 illustrates the sequence of steps followed when undertaking linkage analysis on a large number of samples e.g. performing a genome wide linkage analysis. Steps 2c to 2e are followed if DNA is in short supply. A PEP PCR is performed to increase the amount of DNA available. Steps 3b to 3e are ideally undertaken prior to pooling the whole set so that PCR failures can be identified and PCR can be performed on the failed samples.

The UK genome-wide linkage analysis study -an example of a fluorescent gel

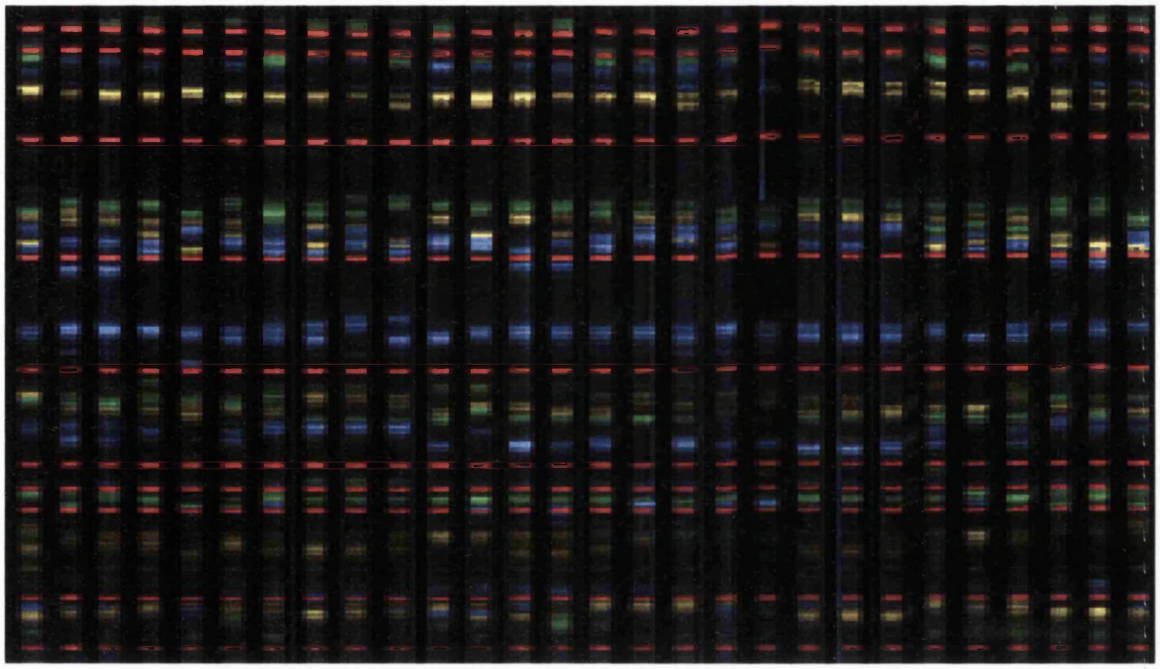


Fig 2.3.2 shows a fluorescent gel from an ABI 373 having been run through the software package Genescan" . The red lines are marker lines allowing the software package Genotyper" to determine the size of each microsatellite allele relative to the known size of the markers. The yellow, blue and green lines are the fluorescently tagged microsatellites. Each column represents one individual and 10 to 20 microsatellite markers per person can be separated by PCR provided the fragments of a similar size are tagged by a different fluorescent colour.

(ARC-UK) National Repository of family material (www.http://arc.man.ac.uk). The affected sibling pairs and their families were recruited following a nation-wide publicity media campaign and by involving clinical rheumatologists throughout the UK in identifying potential families from their practices who were willing to enrol in the study. Multi-centre ethics committee approval was obtained. All family members were examined according to a standard protocol and agreed to a detailed structured interview regarding joint symptoms. A trained metrologist performed a joint examination to detect swelling, tenderness and deformity. Individuals with erosive disease were identified from hand radiographs, reviewed by a rheumatologist, and rheumatoid factor (RF) status was ascertained using a particle agglutination test. Subjects were classified as having rheumatoid arthritis if they satisfied the 1987 American College of Rheumatology (ACR) criteria (Arnett *et al.*, 1988) modified for genetic studies (MacGregor, 1995) (section1.1.2).

2.3.3.2 DNA extraction

As described above for the replication study, DNA from these affected individuals and their first-degree relatives was prepared from peripheral blood. Technicians employed by the ARC-UK National Repository performed all the DNA extraction according to standard techniques. Concentrated DNA was kept in labelled tubes at -70°C.

2.3.3.3 Strategy

Undertaking a linkage analysis of the whole genome in a large number of families involves a considerable quantity of laboratory and statistical work. Three hundred and sixty-eight microsatellite markers were required to span the whole genome with a mean marker distance of ~ 10cM resulting in a minimum of 368 PCR reactions per individual. The ASP families included in the study were those who were the most informative (i.e. most ASPs were

included if DNA was available for themselves, their parents and/or additional siblings). Hence, 650 individuals from 182 families were included in the study meaning that over 250,000 PCR reactions would be required to undertake genotype analysis in these individuals and the resulting PCR products would need to be separated by electrophoresis using over 340 fluorescent gels. Therefore, the laboratory work was carried out in two centres; and identical box set-ups were designed so both centres would use the same numbering and position in the box for each individual. This was to allow both 'quality control' where each centre undertook to genotype the same individuals and also to avoid mistakes during the statistical analysis. Once the box set-ups had been designed (K MacKay, S John) concentrated DNA was diluted with sterile water to 10ng/μl and aliquoted into 16 boxes (eight duplicates for each centre). Statistical analysis was conducted at the Wellcome Trust Centre for Human Genetics in Oxford (K MacKay).

DNA was transported to the WTCHG in Oxford on ice and kept frozen, until required, at -70°C. To allow for the quantity of DNA required to perform the number of PCR reactions needed to genotype each individual a PCR amplification reaction (PEP PCR) was performed on the DNA brought from the ARC-UK National Repository (section 1.3.3, fig 2.3.1). Once amplified, the PEP DNA was aliquoted and frozen at -70°C. Aliquots of the PEP DNA were thawed as required and kept at 4°C whilst the microsatellites were amplified and the gels run (section 2.3.3.7). Each microsatellite was amplified through all boxes of DNA in one go to avoid repetitive freeze-thawing of PCR primers (section 1.3.3).

2.3.3.4 RA affected sibling pairs

Parental DNA was available in 37% of the families and available from additional unaffected siblings in 94% of the remaining families to facilitate the assignment of parental genotypes (table 2.3.1). A total of 252 ASPs were identified from the 182 families included because multiple sibships were

DNA availability	Number of families
Both parents	29
One parent	38
7 siblings	2
5 siblings	4
4 siblings	47
3 siblings	55
2 siblings	7

Table 2.3.1 shows the pedigree structure of the families included in the whole genome screen (WGS). For every affected sibling pair (ASP) additional first-degree relatives were recruited in order to ascertain the identity by descent of each allele studied. This table documents the availability of DNA from parents or additional siblings for each ASP family used in the UK WGS. If parental DNA was unavailable, unaffected siblings were recruited and genotyped if possible allowing the missing parental genotypes to be inferred. In total 650 individuals from 182 families were genotyped.

Affected sibling pairs per family	Number of families
2 ASPs	155
3 ASPs	23
4 ASPs	2
6 ASPs	1

Table 2.3.2 shows the number of ASP families recruited for the WGS and the number of ASPs per family. Some of the families included were made up of multiple sib-ships of between 3 and 8 RA affected siblings and the number of multiple sib-ship families are documented here.

identified in 26 families (table 2.3.2). Of the 650 individuals included in the study, 403 had rheumatoid arthritis, (393 siblings, 10 mothers and 3 fathers). Three of the affected parents were also members of an affected sibling pair. Seventy-six per cent of affected individuals were female, 85% were sero-positive for RF and 80% had developed erosions on hand radiographs. The mean age at disease onset was 39.4 years (SD ± 13.2) and mean disease duration at recruitment was 15.3 years (SD ± 11.3). Sex concordance is summarised in table 2.3.3.

2.3.3.5 HLA-DR status

HLA-DRB1 genotypes were determined using a commercially available semi-automated PCR-SSOP typing technique (Inno-LiPa, Abbott Laboratories Ltd., Maidenhead, UK) (table 2.3.4).

2.3.3.6 Microsatellite markers

Three hundred and sixty-five highly informative microsatellites from the Applied Biosystems (ABI) Prism Linkage Mapping Set Version 2 (LMSv2) marker set (Applied Biosystems, Warrington, UK) were genotyped (appendix 5.4). The microsatellites spanned the whole genome with a mean marker distance of ~ 10cM and a mean heterozygosity of 0.78.

2.3.3.7 Microsatellite genotyping

Semi-automated analysis of microsatellite genotypes was carried out in two centres; chromosomes 1 - 7, 16, 21, 22 were typed at the Wellcome Trust Centre for Human Genetics (WTCHG) and chromosomes 8 - 15, 17 - 20 and X were typed at the Arthritis Research Campaign Epidemiology Unit (ARC-EU). The microsatellites were amplified by PCR using fluorescently labelled primers and reactions were performed in 10μl volumes (containing 50ng of DNA, 400nM each PCR primer, 50μM each dNTP, 0.2 units of Taq polymerase (Bioline, London, UK) in 1-3mM $MgCl_2$ buffer, overlaid with liquid

paraffin). The reactions were performed in 96 well microtitre plates with 30 - 35 cycles of denaturation (1minute, 95°C), primer annealing (30 seconds, 55°C - 60°C) and extension (30 seconds, 72°C). Annealing temperatures and $MgCl_2$ buffer concentration were optimised for each primer. Amplification of each microsatellite was performed separately and the PCR products were then combined into pools of 10 – 20 markers before loading. Products were separated by electrophoresis either on 6% polyacrylamide gels using ABI 373 semi-automated DNA sequencers (Applied Biosystems, Warrington, UK) over three hours (WTCHG) or on 4% polyacrylamide gels and ABI 377 DNA sequencers over two hours (ARC-EU). Products were sized using the program GENESCAN™672 (version 2.1) (Applied Biosystems, Warrington, UK) and genotypes semi automatically assigned using the program GENOTYPER™ (version 1.1.1) (Applied Biosystems, Warrington, UK) (Figs 2.3.2 and 2.3.3). A PCR product from a DNA reference sample (CEPH 1347-02) was included on every gel to monitor possible gel-to-gel variation.

2.3.3.8 Statistical analysis

Error checking:

To minimise data errors extensive error-checking procedures were employed. Allele assignment by GENOTYPER™ was checked manually for all genotypes and the CEPH control sample was used to ensure the consistency of allele assignment. Size data was converted into discrete allele numbers and samples not following Mendelian patterns of inheritance were identified in PEDCHECK (O'Connell and Weeks, 1998) and removed from the data set. The families were then tested with RELATIVE (Goring and Ott, 1997). RELATIVE identifies probable monozygotic twins, half-siblings or unrelated individuals by testing whether the proportion of alleles shared identity by descent at unlinked loci (on the basis of $\geq$ 50 markers) is consistent with the expected proportion for each relative pair.

Sex concordance between affected sibling pairs

Concordance	Total ASPs (n=252)
Female : female ASPs	153 (60.7%)
Male : male ASPs	19 (7.5%)
Female : male ASPs	80 (31.7%)

Table 2.3.3 shows the sex concordance of ASPs included in the UK WGS. As expected for a condition that predominantly affects women there is a predominance of female:female ASPs and only a small percentage of male:male ASPs.

HLA sharing between affected sibling pairs

HLA sharing	Total ASPs (n=252)
Homozygous HLA-DR4	36%
Heterozygous HLA-DR4	48%
No sharing HLA-DR4	16%

Table 2.3.4 shows the HLA sharing of affected sibling pairs (ASPs) included in the UK whole genome screen (WGS). The figures are similar to previous reports showing increased sharing of both HLA-DR4 alleles and a decreased frequency of ASPs with no HLA-DR4 alleles in common. The expected sharing of HLA-DR4 alleles (by chance) would be 25% sharing both, 50% sharing one, 25% sharing neither HLA-DR4 allele.

Inter and intra-observer variation:

To assess inter-observer variation, a panel of samples was genotyped for six microsatellites on chromosome 7 in both centres. A further ten markers were re-genotyped in a blinded fashion by the researchers who were undertaking the genotype analysis (K MacKay, S Eyre, A Barton, A Myerscough). Intra-observer variation was then determined.

Non-parametric analysis:

Allele frequencies were calculated from all of the scored genotypes using the software packages DOWNFREQ (J. Terwilliger, unpublished) (WTCHG) or SPLINK (version 1.05, David Clayton, MRC Biostatistics Unit, Cambridge) (ARC-EU). Sibling pair analysis methods were then used to test for excess allele sharing between affected siblings. Non-parametric single point and multipoint analysis was implemented in MAPMAKER/SIBS (version 2) (Kruglyak and Lander, 1995). Multiple sibships were given a conservative weighting of 2/n to account for their lack of independence (Meunier *et al.*, 1997). Analysis of markers on the X chromosome was carried out using MAPMAKER/SIBS modified for X-linked loci (Cordell *et al.*, 1995). The program calculates sharing between sister-sister, brother-brother and sister-brother affected sib pairs.

Both multipoint and single point analyses have been presented. Suggested thresholds of LOD ≥ 3.6 ($p \leq 2 \times 10^{-5}$) for significant linkage and LOD ≥ 2.2 ($p \leq 7.4 \times 10^{-4}$) for suggestive linkage were used (Lander and Kruglyak, 1995). Nominal evidence of linkage was taken to be LOD ≥ 0.8 ($p < 0.05$) by either multipoint or single point analysis.

Summary of positive single or multipoint results from the UK whole genome screen

Marker	Chromo-some	Distance from Telomere (cM)	Single point (p)	Multipoint (p)
D1S2842	1q	277cM		0.05
D1S2836	1q	290cM		0.05
D3S2338	3p	42cM	0.05	
D4S1592	4q	69.5cM	0.05	
D6S276	6q	1cM from DRB1		0.000003
DRB1	6q	DRB1		LOD = 1.9
D6S434	6q	109cM	0.0006	0.0007
D7S484	7p	53.5cM	0.05	
D10S192	10q	124cM	0.05	
D10S217	10q	158cM	0.05	
D14S283	14q	28cM		0.05
D14S275	14q	28cM	0.05	0.05
D14S276	14q	56cM	0.05	0.05
D16S3103	16p	32cM	0.05	
D21S1256	21q	9.7cM	0.05	
DXS1106	Xq	66.6cM	0.05	

Table 2.3.5 summarises the positive results identified by the UK whole genome screen (WGS). The microsatellite markers are listed along with the chromosomal arm they can be found on and the distance of the marker in -cM from the telomere. Both single and multipoint results are listed.

Marker positions obtained from Marshfield map (sex-averaged positions)
http:\\www.research.marshfieldclinic.org/genetics
cM = centimorgans

2.3.4 Results

2.3.4.1 Linkage analysis:

The results of the non-parametric, single point and multipoint analysis implemented in MAPMAKER/SIBS are shown in figures 2.3.2-2.3.24. Positive results identified by single or multipoint analysis are summarised in table 2.3.5. Linkage to the HLA region on chromosome 6 was confirmed by multipoint analysis (max LOD = 4.7 at marker D6S276, p=0.000003, 1cM from HLA-DRB1). The allele-sharing ratio for DRB1 was 16:49:35 for zero, one and two allele sharers. This significant increase in the inheritance of 2 alleles identical by descent provides evidence for linkage to DRB1 (single-point LOD score 1.9). Suggestive evidence of linkage by multipoint analysis was identified on chromosome 6q (D6S434, p=0.0007) at 109-cM. Nominal evidence of linkage by multipoint analysis was found for 3 additional linkage intervals: 1q (D1S2842 and D1S2836, p=0.05) at 277-290cM, 14q (D14S283, p=0.05 and D14S275, p=0.03) at 28cM and 14q (D14S276, p=0.03) at 56.4cM. Single point analysis identified suggestive linkage for one marker on chromosome 6q (D6S434, p=0.0006) at 109-cM. Nominal linkage ($p<0.05$) was identified by single point analysis for 10 additional linkage intervals on chromosomes 3p (D3S2338 at 42cM), 4q (D4S1592 at 69.5cM), 7p (D7S484 at 53.5cM), 10q (D10S192 at 124cM), 10q (D10S217 at 158cM), 14q(D14S275 at 28cM), 14q(D14S276 at 56.4cM), 16p (D16S3103 at 32cM), 21q (D21S1256 at 9.7cM) and Xq (DXS1106 at 66.6cM). Comparisons are made with data from the European and North American whole genome screens in Tables 2.3.7.

2.3.4.2 Microsatellite markers:

The microsatellite markers spanned the genome with a mean marker distance of 10cM and a median of 9.3cM. Eighty-eight per cent of markers were within 14cM of each other and only 11 markers had a gap of > 20cM,

PCR success per chromosome

Chromo-some	PCR success (%)	Centre	Chromo-some	PCR success (%)	Centre
1	92	WTCHG	13	88	ARCERU
2	91	WTCHG	14	88	ARCERU
3	92	WTCHG	15	79	ARCERU
4	89	WTCHG	16	95	WTCHG
5	90	WTCHG	17	82	ARCERU
6	96	WTCHG	18	79	ARCERU
7	92	WTCHG	19	68	ARCERU
8	75	ARCERU	20	78	ARCERU
9	85	ARCERU	21	99	WTCHG
10	81	ARCERU	22	92	WTCHG
11	84	ARCERU	X	77	ARCERU
12	86	ARCERU			

WTCHG = Wellcome trust centre for Human Genetics
ARCEU = Arthritis Research Campaign Epidemiology Research Unit

Table 2.3.6 summarises the overall PCR success per chromosome for each microsatellite marker amplified. This is presented as an overall percentage and refers to the percentage of DNA samples per chromosome where all the microsatellites were successfully amplified.

using the Marshfield map (table 2.3.6), as a result of marker failure (table 5.4.1). The mean heterozygosity was 0.78±0.07 and the mean polymorphism information content (PIC) was 0.75±0.08. Tables 2.3.5 and 5.2.1) summarise the percentage PCR success for each marker used in the WGS.

2.3.4.3 Error checking:

Fifteen individuals not fitting Mendelian inheritance patterns were identified by PEDCHECK (O'Connell and Weeks, 1998) and removed from the analysis. Three previously unrecognised half-siblings and two further individual samples not fitting Mendelian inheritance patterns were identified by RELATIVE (Goring and Ott, 1997) and also removed from the analysis.

2.3.4.4 Inter and intra-observer variation:

There was 98.9 per cent concordance for the six microsatellite markers on chromosome 7 that were genotyped at both centres. Intra-observer variation, measured by re-genotyping 10 markers blindly, ranged from 0 - 3%.

2.3.5 Discussion

In keeping with two previously reported RA whole genome screens (Cornelis *et al.*, 1998; Jawaheer *et al.*, 2001) this study confirmed genetic linkage to the HLA region on chromosome 6. In addition, suggestive evidence of linkage ($p< 7.4 \times 10^{-4}$) to a non-HLA region on 6q was identified by both single and multi-point analysis. Nominal evidence of linkage ($p<0.05$) to 10 other regions was detected by single point analysis (chromosomes 3p, 4q, 7p, 2 regions on 10q, two regions on 14q, 16p, 21q and Xq) and to three linkage intervals by multipoint analysis (chromosomes 1q, and 2 regions on 14q).

Summary of potential RA loci identified by three independent whole genome screens in Europe, USA and UK

Chromosomal Location	UK	ECRAF	NARAC
1q	D1S238 (203cM) p=0.05 D1S2842 (277cM) p=0.05 D1S2836 (290cM) p=0.05	D1S238 (203cM) (SP)p=0.022	D1S235 (259cM) p=0.005
3q		D3S1267(139cM) (IDDM9) p=0.039	D3S4523 (138cM) (IDDM9) p=0.029
6q	D6S434 (109cM) p=0.0007		D6S1021 (112.2cM) p=0.008
8p		D8S1825 (15.4cM) (SP) p=0.040	D8S277 (8.3cM) p=0.009
12q		D12S95 (96.1cM) p=0.043	D12S1052 (83.2cM) p=0.023
14q	D14S275 (28cM) p=0.03		D14S1280 (25.9cM) p=0.017
14q	D14S276 (56.4cM) p=0.05	D14S285 (59.4cM) (SP) p=0.049	D14S587 (55.8cM) p=0.037
16p		D16S420 (44.5cM) p=0.039	D16S403 (43.9cM) p=0.004
18q		D18S474 (71.3cM) p=0.012	D18s858 (80.4cM) p=0.043

Table 2.3.7 compares the overlapping regions of positive linkage from the three major whole genome screens (WGS) undertaken in European, North American and UK populations with RA.
(SP= single-point, all other results = multi-point.).
Marker positions obtained from Marshfield map (sex-averaged positions)
http:\\www.research.marshfieldclinic.org/genetics
cM = centimorgans

Regions identified as nominal linkage intervals by two or more WGSs are of greater interest than sites identified by only one study, especially where corroborative evidence from other autoimmune diseases or animal studies is available. Six of the non-HLA regions identified in this study were also reported by the ECRAF (Cornelis *et al.*, 1998) and/or NARAC [Jawaheer, 2001 #1031] studies (chromosomes 3q, 6q, 10q, 14p, 14q and 16p) (table 2.3.1) and a number overlap with regions linked to other autoimmune diseases (Tables 2.3.8). The non-HLA region on 6q with suggestive evidence of linkage was also identified by the NARAC study (Jawaheer *et al.*, 2001). On chromosome 10, the linkage interval including D10S185 (116cM) and D10S192 (124cM) was identified in the NARAC study (Jawaheer *et al.*, 2001) and has also been identified in an ankylosing spondylitis WGS (Brown *et al.*, 1998; Laval *et al.*, 2001). In addition, the interval around D10S217 (167cM) has been identified in a recent second-generation WGS in insulin dependent diabetes (IDDM) (Concannon *et al.*, 1998). The region on 14q including D14S276 (56cM) was identified by both the ECRAF (Cornelis *et al.*, 1998) and NARAC (Jawaheer *et al.*, 2001) studies as was the interval on 16p (D16S3103). Recently, a WGS conducted using mice with proteoglycan-induced arthritis (Otto *et al.*, 1999) also identified a linkage interval (named p*gia*-3) homologous to this same region on 16p.

In keeping with the NARAC study (Jawaheer *et al.*, 2001) no supporting evidence was found for the IDDM-6 locus on chromosomes 18, a site which had shown evidence for linkage in the ECRAF study (Cornelis *et al.*, 1998). Previous investigations using some of the same RA ASP families as in this study have reported linkage to 3 candidate genes (MacKay *et al.*, 2000). The current study did not attempt to replicate these findings. Using a number of single nucleotide polymorphism and microsatellite markers spanning the Interleukin 1 (IL-1) gene cluster, Cox *et al.* detected some evidence of linkage using the combined Transmission disequilibrium test (TDT) and sib-TDT in a

subset of families sharing only one or zero DRB1 alleles identical-by-descent. However, they found no evidence of linkage using non-parametric allele sharing methods (Cox *et al.*, 1999a). John *et al.* reported linkage to a single marker mapping to an intron within the Estrogen Synthase gene (CYP-19) (John *et al.*, 1999). In the current study the closest marker mapped 5cM from CYP-19 and this may explain the failure to replicate this finding. Fife *et al.* investigated Corticotrophin Releasing Hormone (CRH) as a candidate RA susceptibility locus (Fife *et al.*, 2002; Fife *et al.*, 2000). Linkage to a 10cM region adjacent to CRH was reported. No statistically significant evidence of linkage was detected in the current study although there was some overlap in the families used in the two investigations. It is interesting to note, however, that there is an apparent small peak on multi-point linkage analysis to the same region in the current study (Figure 2.3.8). With the marker density used in this study linkage to these (and other postulated disease) loci could have been missed and they cannot, therefore, be excluded as potential disease genes.

Power calculations suggest that the current study should have 80% power to detect loci contributing $\lambda = 1.6$ using the threshold for detection as a LOD $\geq$ 1.0. This, in fact, may be an overestimate because power calculations typically assume that all families are fully informative (*i.e.* DNA available from all parents and 100% marker heterozygosity) and all affected sibling pairs are independent (which is not the case in families with more than two affected siblings). Recruitment of complete affected sibling pair families is difficult in a late onset disease such as rheumatoid arthritis. Thus, parental DNA was available in just over one third of our families. However, one of the strengths of this study is that DNA from unaffected siblings was available in 94% of the remaining families and this was used to infer parental haplotypes when these were not directly available.

Comparing loci detected by the UK WGS with the results of other linkage analyses of RA and related autoimmune conditions.

Chromosomal Location	UK Marker (map position)	ECRAF Marker (map position)	NARAC Marker (map position)	Studies of related diseases with linkage to the region
3	D3S2338 (42cM)		D3S3038 (44.8cM)	
4	D4S1592 (69.5cM)			
7	D7S484 (53.5cM)			D7S484 MS + CD/UC + asthma[a]
10	D10S185 (116.3cM) D10S192 (124cM) D10S217 (158cM)		D10S2470 (112.6cM)	D10S217 IDDM[a]
16	D16S3103 (32cM)	D16S420 (42.5cM) D16S401 (47cM)	D16S403 (43.9cM)	Pgia3[b] IBD 1[c]
21	D21S1256 (9.7cM)			
X	DXS1106 (66.6cM)			

[a]: MS = multiple sclerosis; CD/UC = Crohn's disease/ulcerative colitis; IDDM = insulin dependent diabetes mellitus. Linkage results summarised in ref Becker 98
[b]: Ref [Brown, 1998; Laval, 2001]
[c]: Ref [Concannon, 1998]
Marker locations obtained from Marshfield website:
http:\\www.research.marshfieldclinic.org/genetics

Table 2.3.8 compares the loci detected by single point linkage analysis of the UK whole genome screen (WGS) with the results of other linkage analyses of affected sibling pairs (ASPs) with RA or related autoimmune conditions. Regions identified as nominal linkage intervals by two or more WGS are of greater interest than are sites identified by only one study, especially when corroborative evidence from other autoimmune disease or animal studies is available.

Because of the high chance of false positives in a genome-wide linkage study the stringent thresholds for significance ($p<2.2\times10^{-5}$) recommended by Lander and Kruglyak have been applied (Lander and Kruglyak, 1995). However, if stringent thresholds for significance are interpreted too strictly in a disease such as RA with a relatively low λ_s it would be easy to overlook potentially relevant evidence emerging from WGSs. In order to make best use of the information arising from this and other similar sized WGS, a number of strategies are available. One proposal is to perform the initial screen in a relatively small data-set and then to test those loci demonstrating nominal evidence of linkage ($p<0.05$) in a larger data-set (Holmans and Craddock, 1997). In this respect, the study could be seen as a replication data set in which to test positive linkages detected in the US and European WGS. As outlined above, a number of positive linkages from the NARAC and ECRAF studies have also shown evidence of linkage in our data set suggesting that true RA susceptibility genes may map to these regions. A complementary but alternative approach would be to undertake a meta-analysis analysis of the genotype data already obtained by published whole genome screens. By combining this data it should be possible to make more robust interpretations of the evidence so focusing the efforts of the replication studies.

To conclude, the UK whole genome linkage analysis of RA susceptibility loci confirmed significant evidence of linkage in the HLA region by single point and multi-point analysis. Suggestive linkage was identified by single and multipoint analysis at a non-HLA region on 6q and nominal evidence of linkage ($p<0.05$) has been found for 10 other regions by single point analysis on chromosomes 3p, 4p, 7p, 2 regions of 10q, 2 regions of 14q, 16p, 21p and Xq and 3 non-MHC regions by multipoint analysis on chromosomes 1q and 2 regions on 14q.

Results of single and multi-point linkage analysis for chromosome 1

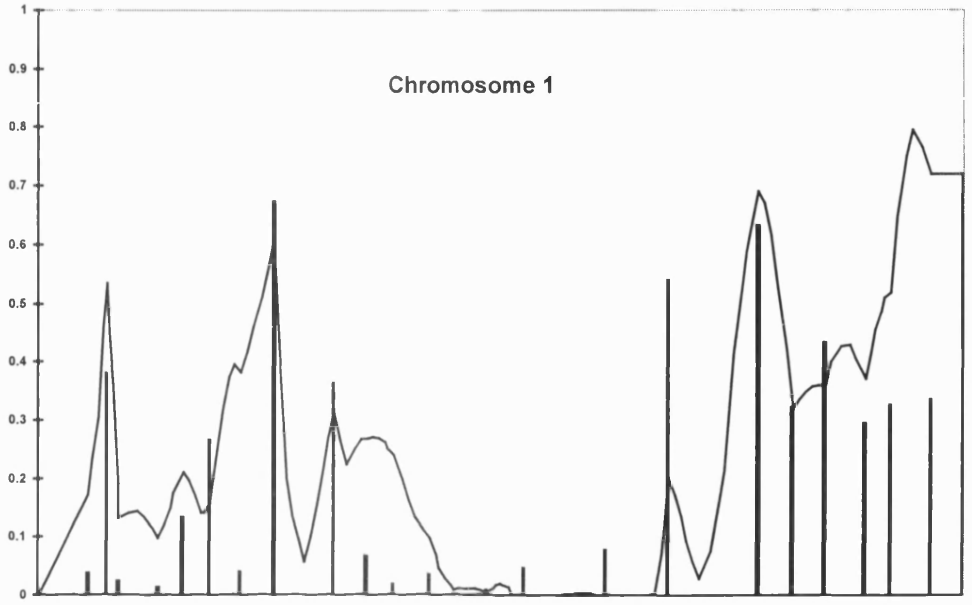


Fig 2.3.3 shows the results of single and multi-point linkage analysis for chromosome 1 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 1. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 1.

Results of single and multi-point linkage analysis for chromosome 2

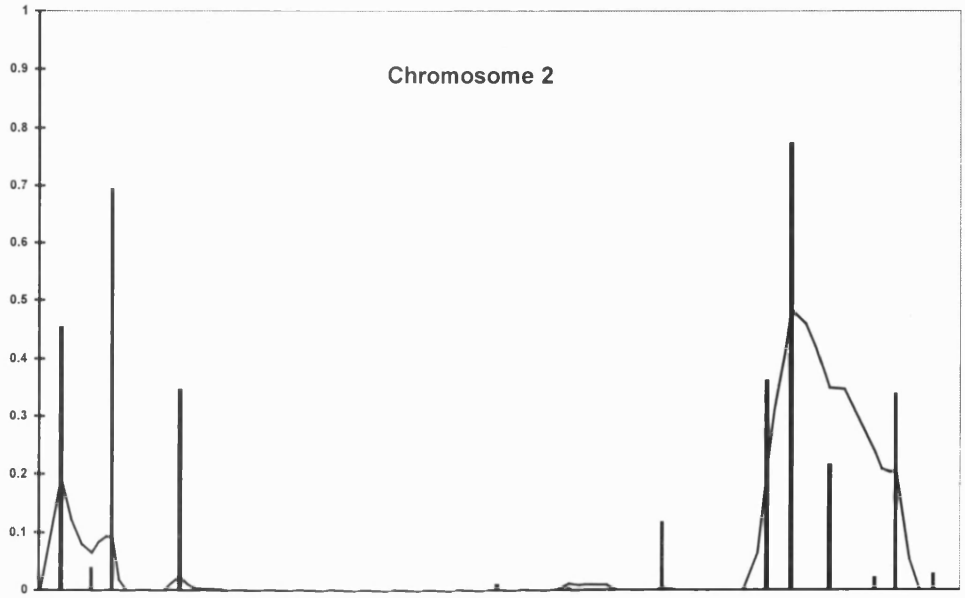


Fig 2.3.4 shows the results of single and multi-point linkage analysis for chromosome 2 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 2. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 2

Results of single and multi-point linkage analysis for chromosome 3

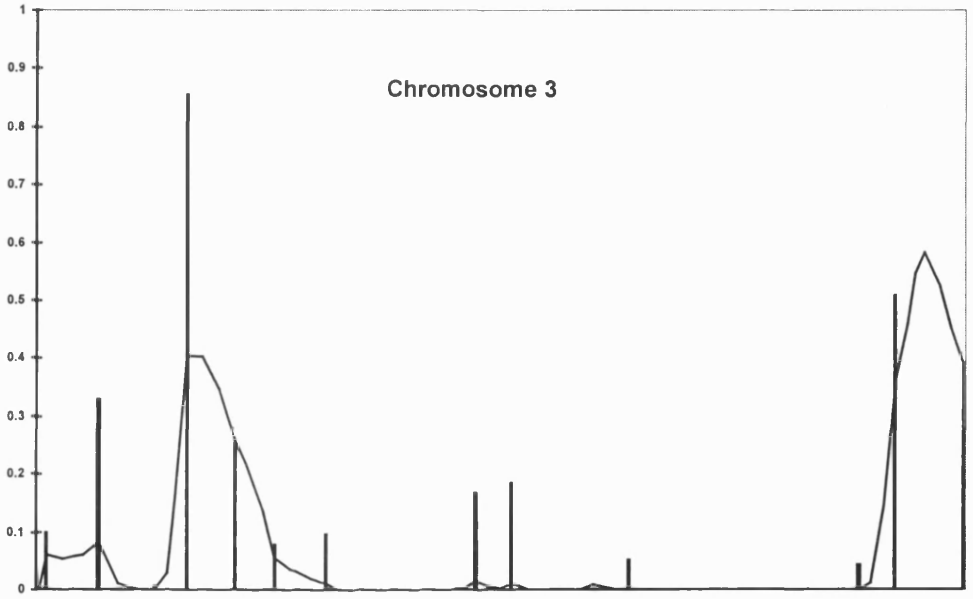


Fig 2.3.5 shows the results of single and multi-point linkage analysis for chromosome 3 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 3. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 3.

Results of single and multi-point linkage analysis for chromosome 4

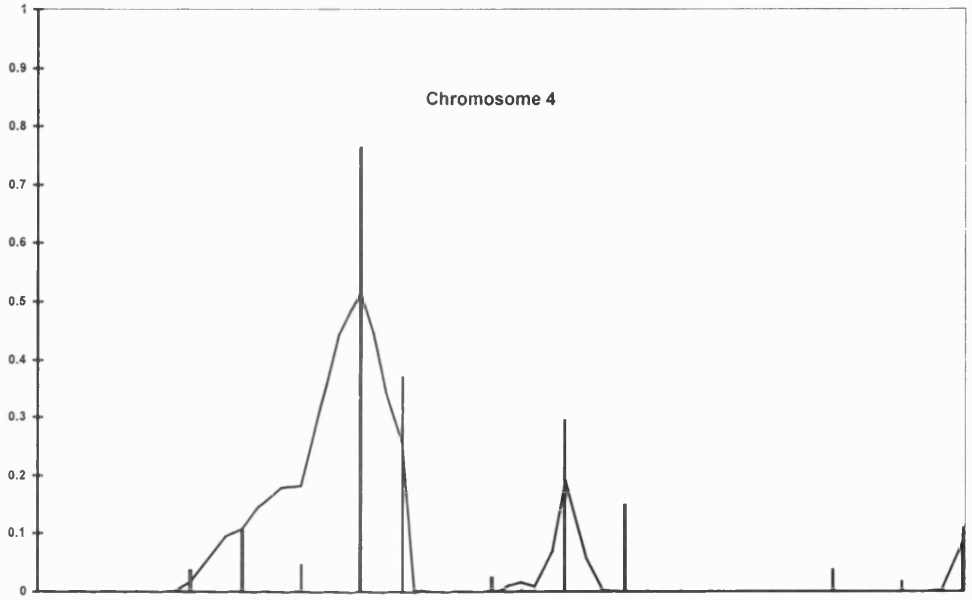


Fig 2.3.6 shows the results of single and multi-point linkage analysis for chromosome 4 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 4. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 4.

Results of single and multi-point linkage analysis for chromosome 5

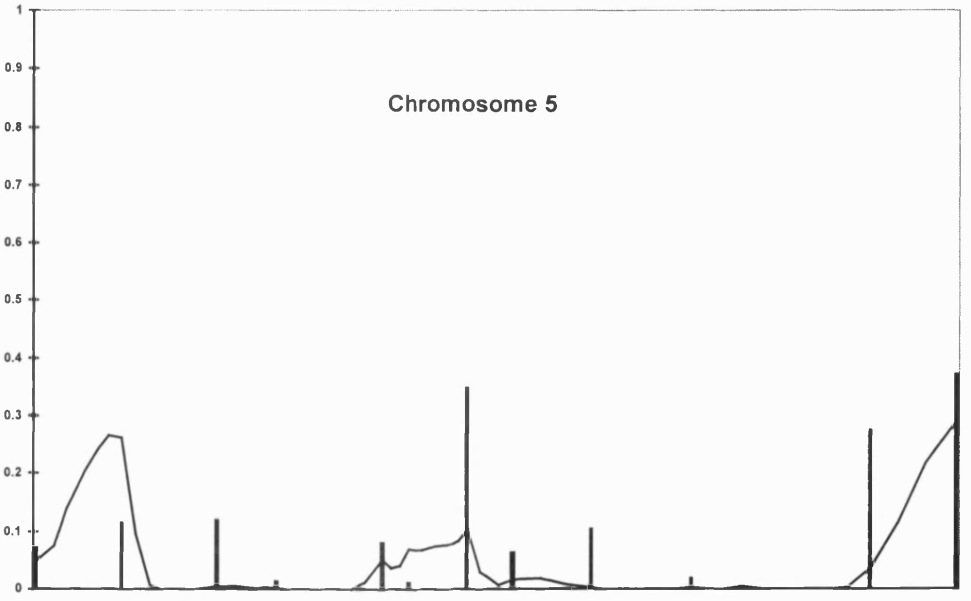


Fig 2.3.7 shows the results of single and multi-point linkage analysis for chromosome 5 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 5. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 5.

Results of single and multi-point linkage analysis for chromosome 6

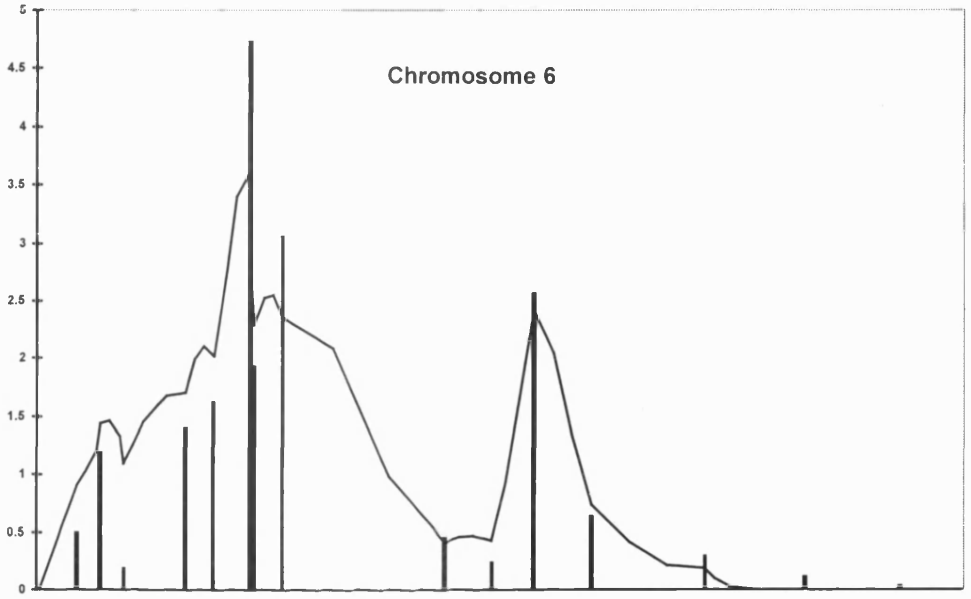


Fig 2.3.8 shows the results of single and multi-point linkage analysis for chromosome 6 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 6. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 6.

Results of single and multi-point linkage analysis for chromosome 7

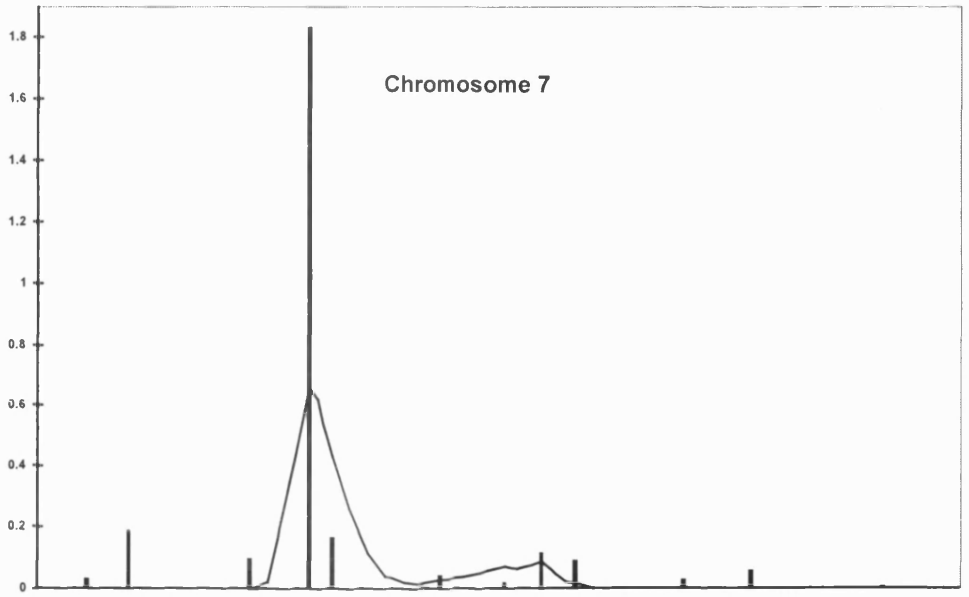


Fig 2.3.9 shows the results of single and multi-point linkage analysis for chromosome 7 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 7. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 7.

Results of single and multi-point linkage analysis for chromosome 8

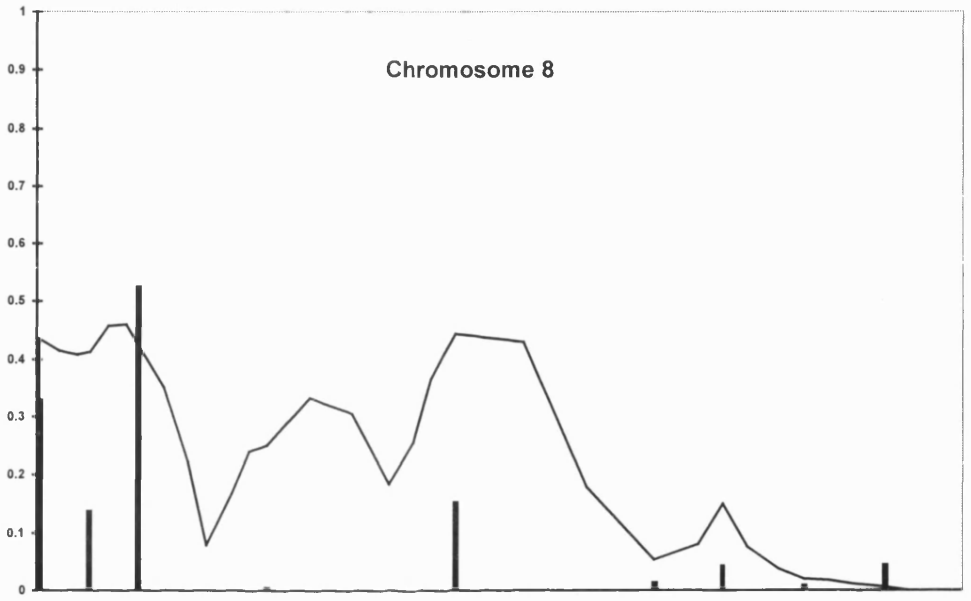


Fig 2.3.10 shows the results of single and multi-point linkage analysis for chromosome 8 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 8. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 8.

Results of single and multi-point linkage analysis for chromosome 9

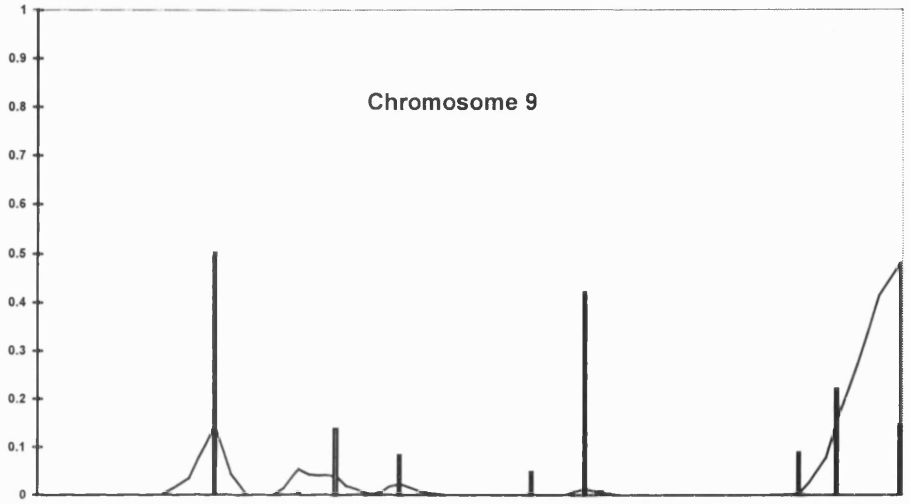


Fig 2.3.11 shows the results of single and multi-point linkage analysis for chromosome 9 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 9. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 9.

Results of single and multi-point linkage analysis for chromosome 10

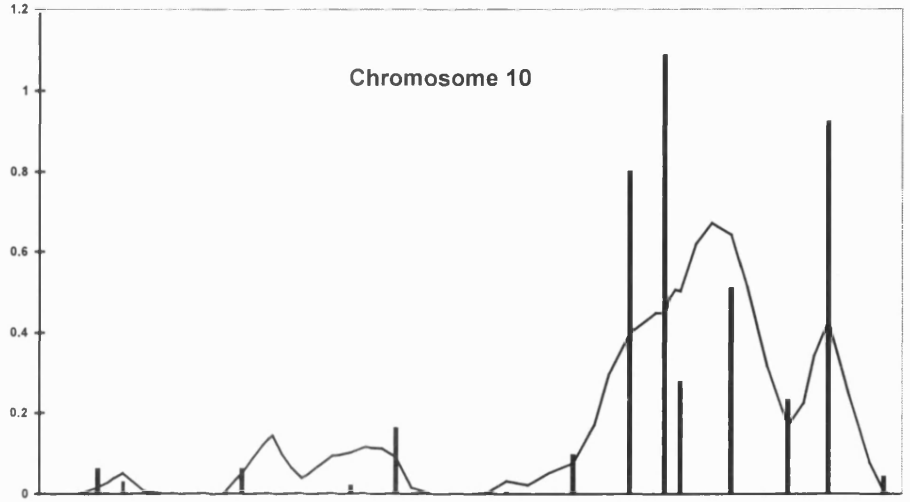


Fig 2.3.12 shows the results of single and multi-point linkage analysis for chromosome 10 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 10. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 10.

Results of single and multi-point linkage analysis for chromosome 11

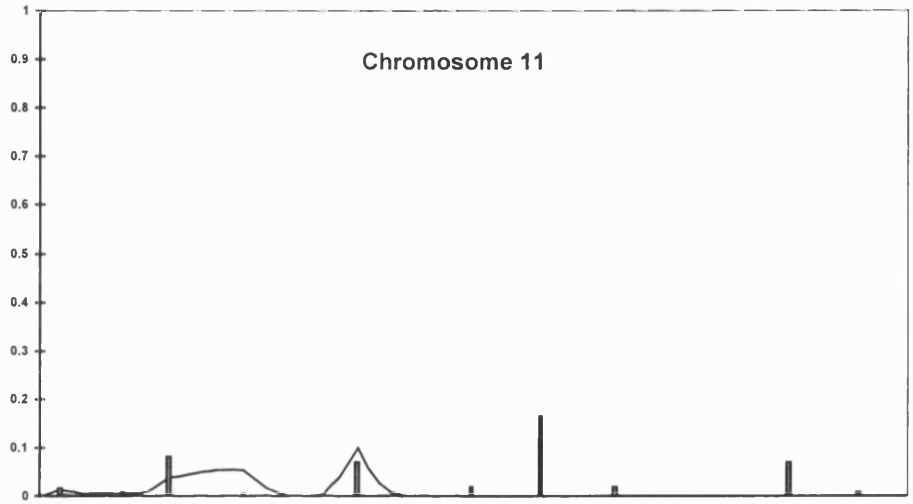


Fig 2.3.13 shows the results of single and multi-point linkage analysis for chromosome 11 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 11. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 11.

Results of single and multi-point linkage analysis for chromosome 12

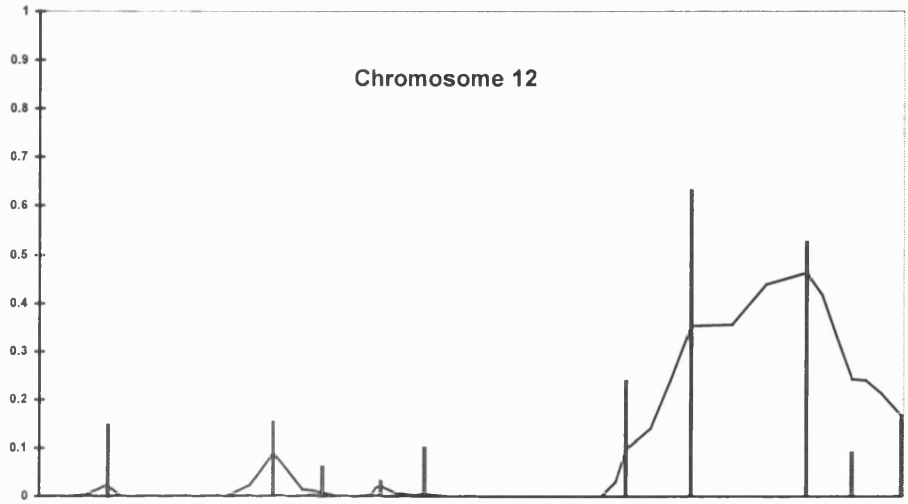


Fig 2.3.14 shows the results of single and multi-point linkage analysis for chromosome 12 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellie markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 12. The distance between them on the graph gives some indication of the genetic distance between them on chromosome12.

Results of single and multi-point linkage analysis for chromosome 13

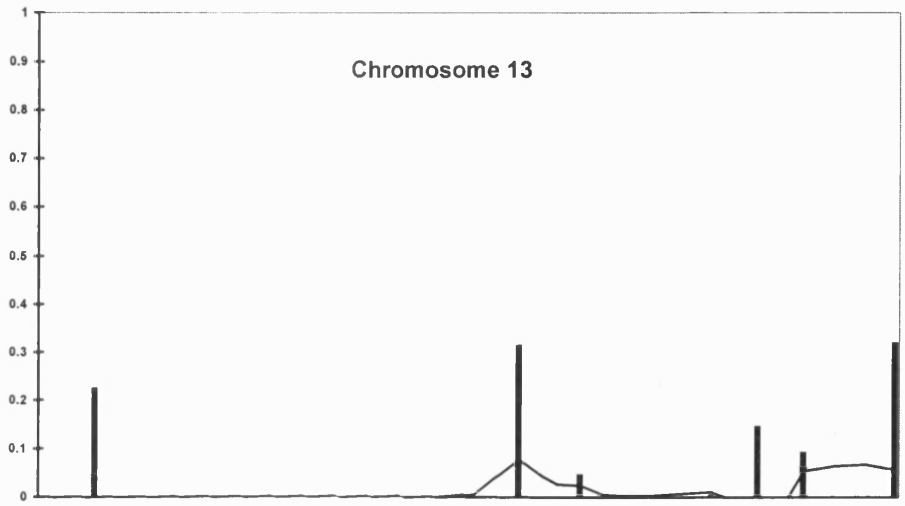


Fig 2.3.15 shows the results of single and multi-point linkage analysis for chromosome 13 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 13. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 13.

Results of single and multi-point linkage analysis for chromosome 14

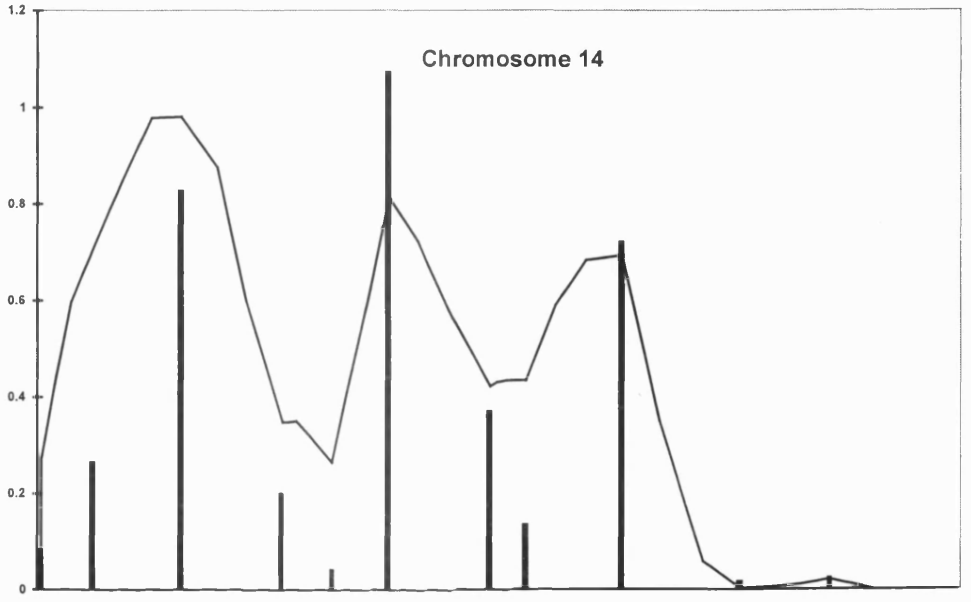


Fig 2.3.16 shows the results of single and multi-point linkage analysis for chromosome 14 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 14. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 14.

Results of single and multi-point linkage analysis for chromosome 15

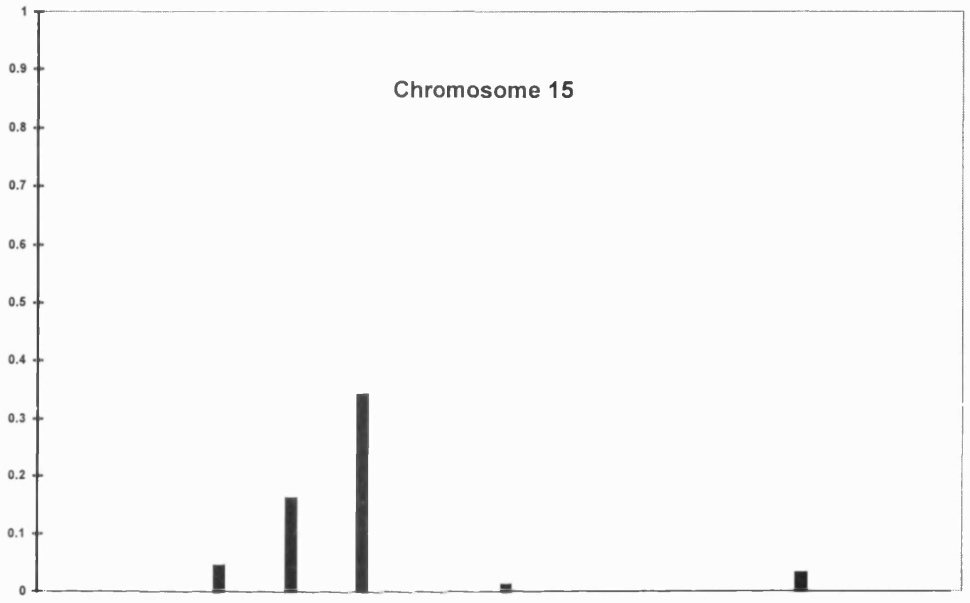


Fig 2.3.17 shows the results of single and multi-point linkage analysis for chromosome 15 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 15. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 15.

Results of single and multi-point linkage analysis for chromosome 16

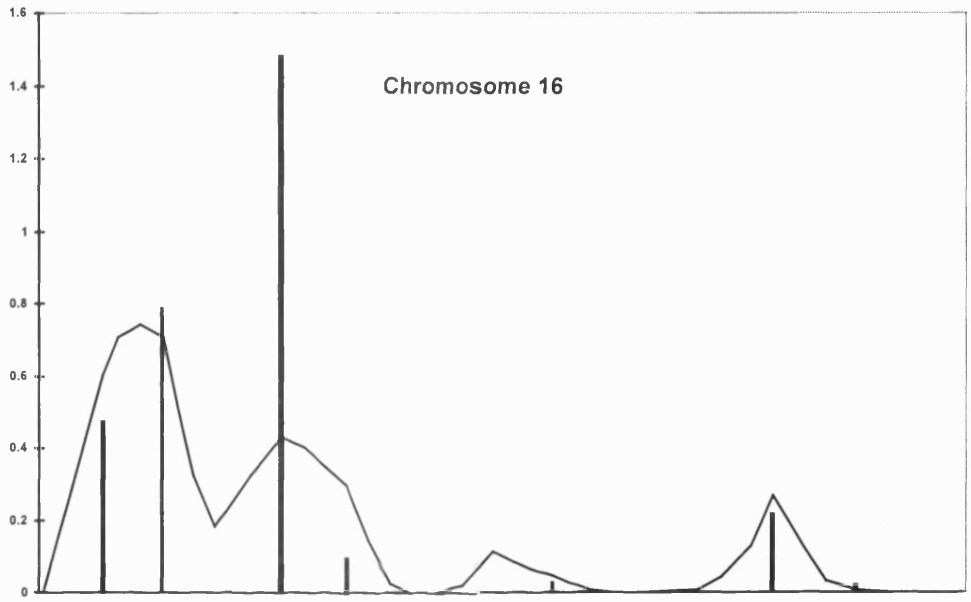


Fig 2.3.18 shows the results of single and multi-point linkage analysis for chromosome 16 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 16. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 16.

Results of single and multi-point linkage analysis for chromosome 17

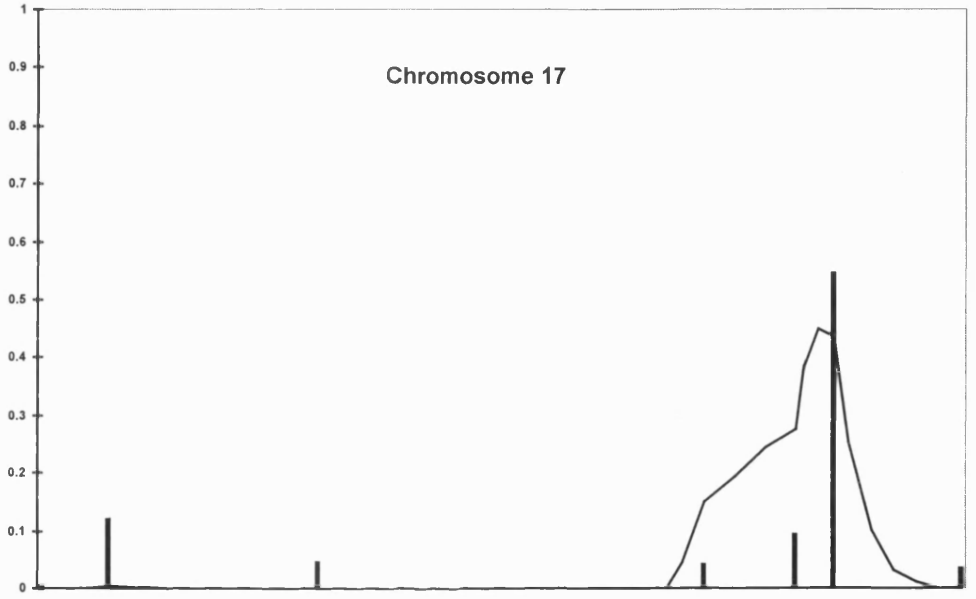


Fig 2.3.19 shows the results of single and multi-point linkage analysis for chromosome 17 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 17. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 17

Results of single and multi-point linkage analysis for chromosome 18

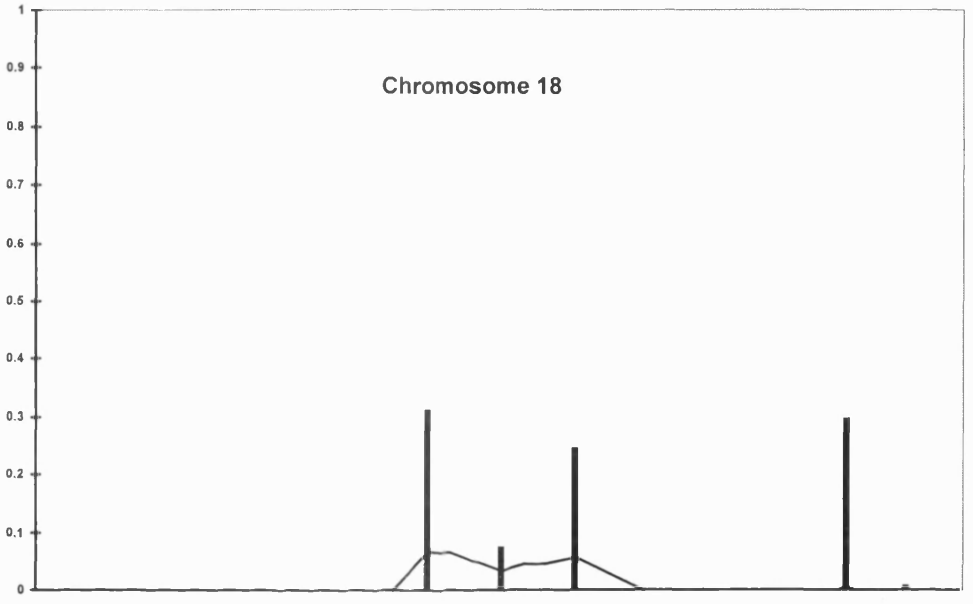


Fig 2.3.20 shows the results of single and multi-point linkage analysis for chromosome 18 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 18. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 18.

Results of single and multi-point linkage analysis for chromosome 19

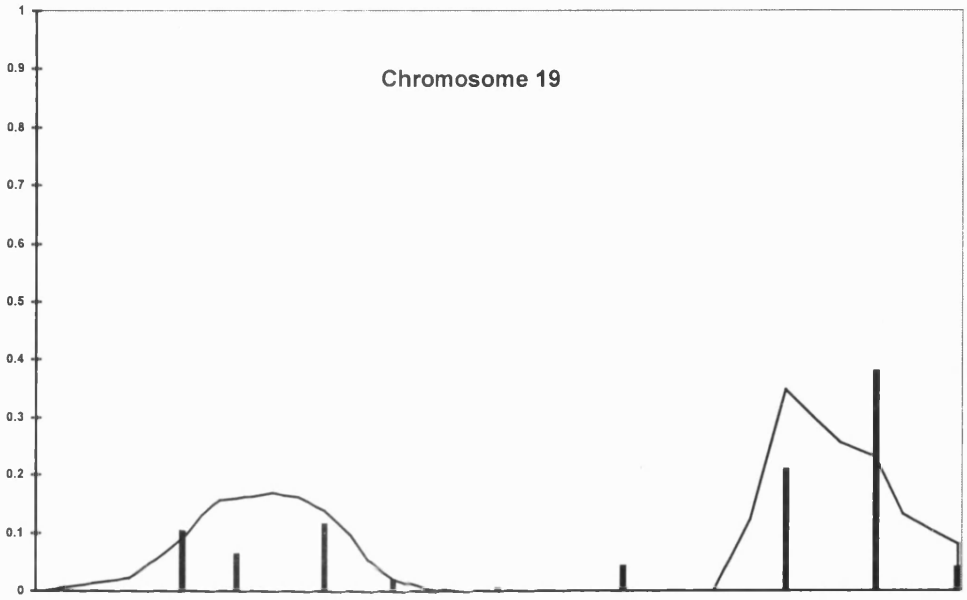


Fig 2.3.21 shows the results of single and multi-point linkage analysis for chromosome 19 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 19. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 19.

Results of single and multi-point linkage analysis for chromosome 20

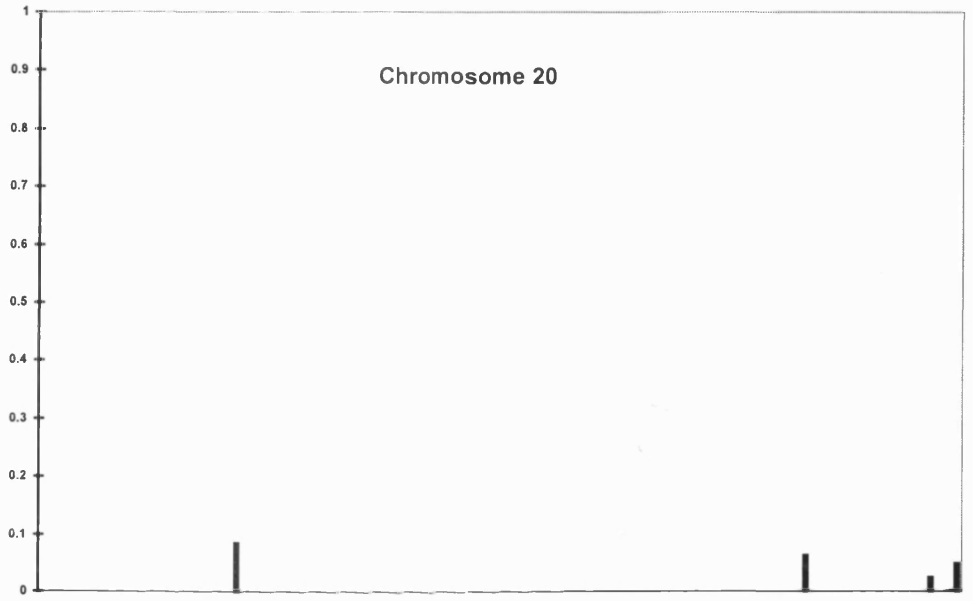


Fig 2.3.22 shows the results of single and multi-point linkage analysis for chromosome 20 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 20. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 20.

Results of single and multi-point linkage analysis for chromosome 21

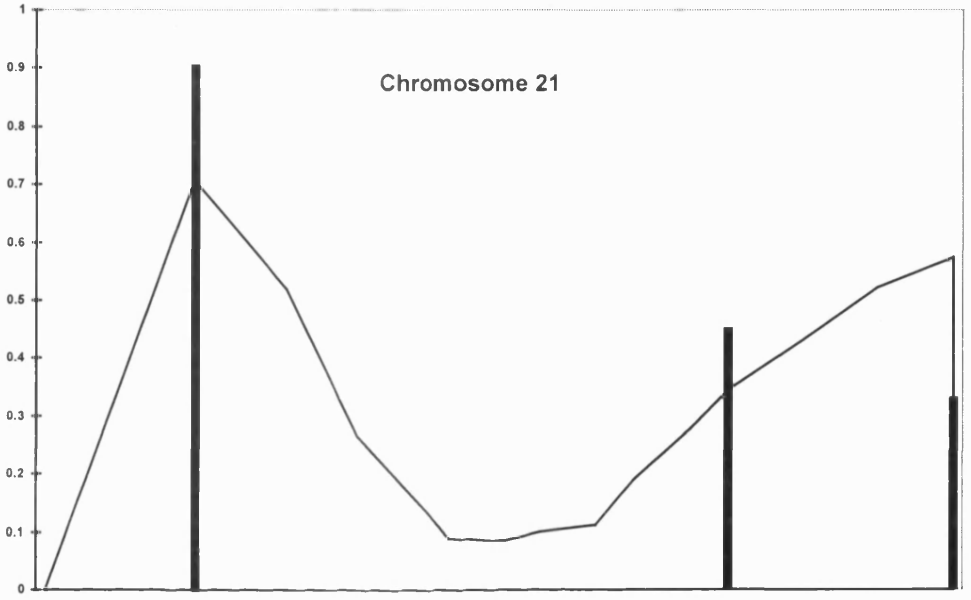


Fig 2.3.23 shows the results of single and multi-point linkage analysis for chromosome 21 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 21. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 21.

Results of single and multi-point linkage analysis for chromosome 22

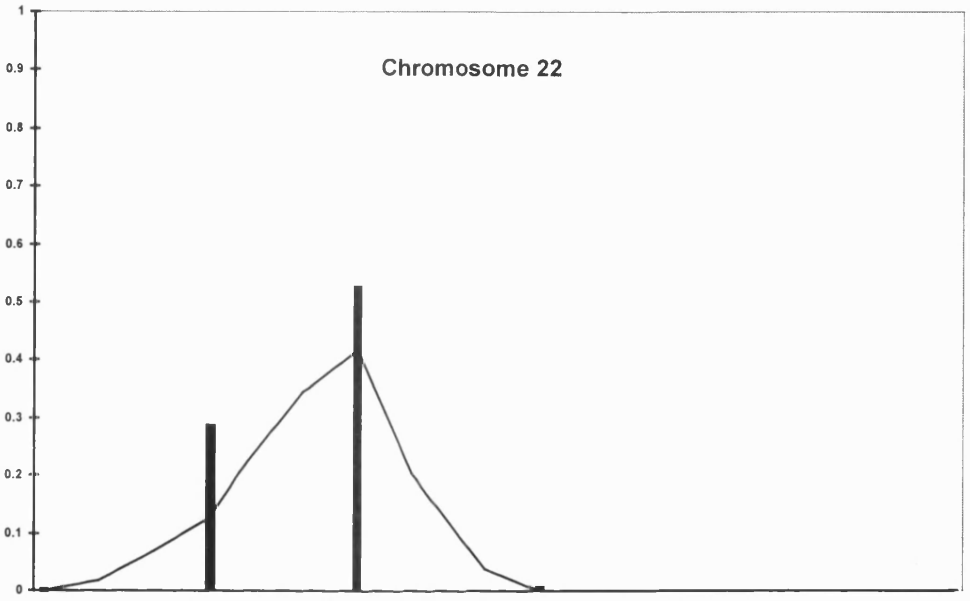


Fig 2.3.24 shows the results of single and multi-point linkage analysis for chromosome 22 using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on chromosome 22. The distance between them on the graph gives some indication of the genetic distance between them on chromosome 22.

Results of single and multi-point linkage analysis for X chromosome

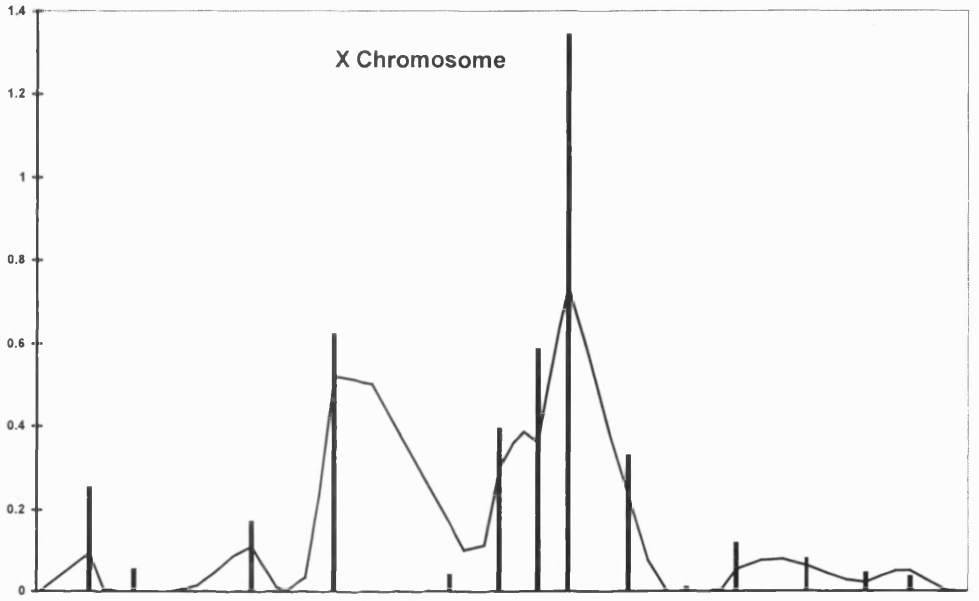


Fig 2.3.25 shows the results of single and multi-point linkage analysis for X chromosome using MAPMAKER/SIBS. The bars represent the results of single point linkage analysis and the continuous line, the results of multi-point linkage analysis. All the microsatellite markers used for the analysis are documented along the X axis in the order that they are to be found on X chromosome. The distance between them on the graph gives some indication of the genetic distance between them on X chromosome.

Chapter 3 – section 1

CHAPTER 3: CANDIDATE GENE STUDIES

Rheumatoid arthritis susceptibility and interleukin 10 (IL10)

3.1 Introduction

Although linkage analysis is an integral part of the search for susceptibility genes it is both expensive and time consuming. The search for candidate genes, undertaken in parallel, can be a useful method of investigating biologically plausible candidate genes at considerably less cost in time and resources. Genes that control cytokine expression are among the most obvious candidates to study in a chronic inflammatory disease such as rheumatoid arthritis. The reasons why interleukin 10 (IL10) may be involved in susceptibility and/or severity to RA are set out below.

3.1.1 Immunology

Pro inflammatory cytokines such as tumor necrosis factor (TNF) and interleukin 1 (IL-1) play a significant pathogenic role in the chronic synovitis and progressive joint destruction of rheumatoid arthritis (Chu *et al.*, 1992). Recently it has been shown that RA can be ameliorated by treatments specifically targeting the activity of TNF (Lipsky *et al.*, 2000) or IL-1 (Arend and Guthridge, 2000). Anti-inflammatory cytokines can also be found in affected joints (Chomarat *et al.*, 1995; Katsikis *et al.*, 1994) where they can inhibit many pro-inflammatory cytokines (de Waal Malefyt *et al.*, 1991; Fiorentino *et al.*, 1991a; Fiorentino *et al.*, 1991b). It has been postulated that

chronic synovitis may reflect an imbalance in pro- and anti-inflammatory cytokine production in rheumatoid arthritis (Weckmann and Alcocer-Varela, 1996). Consequently, the potential beneficial effects of anti-inflammatory cytokines such as interleukin 10 (IL10) (Keystone *et al.*, 1998) and interleukin 4 (IL-4) (Woods *et al.*, 1999) in RA are of great interest.

3.1.2 Interleukin 10, an immuno-modulatory cytokine

IL10 is produced by a variety of cell types, including monocytes (de Waal Malefyt *et al.*, 1991), B lymphocytes (Burdin *et al.*, 1993), activated $CD4^+$ and $CD8^+$ T cells (Spits and de Waal Malefyt, 1992) and mast cells (Ishizuka *et al.*, 1999). IL10 is a potent up-regulator of B cell production and differentiation (Rousset *et al.*, 1992) but has anti-inflammatory capabilities that can directly down-regulate TNFα, IL-1, IL-8 and IFNγ production (Cassatella *et al.*, 1993; de Waal Malefyt *et al.*, 1991; Fiorentino *et al.*, 1989; Fiorentino *et al.*, 1991a; Katsikis *et al.*, 1994). By reducing HLA class II expression on macrophages IL10 can also inhibit antigen presentation (Mottonen *et al.*, 1998). For example, synovial fluid macrophages from patients with rheumatoid arthritis incubated with IL10 had a significantly reduced capacity to activate T cells in a mixed lymphocyte reaction (Mottonen *et al.*, 1998). Also, pulsed methylprednisolone used to produce a rapid and sustained decrease in rheumatoid arthritis disease activity also leads to a sustained increase in IL10 production by peripheral blood mononuclear cells for at least 6 weeks (Verhoef *et al.*, 1999).

Transgenic mice with inactivated IL10 show normal development of lymphocytes and antibody responses but most animals are growth retarded, anaemic and suffer from chronic enterocolitis. Enhanced epithelial MHC class II expression and uncontrolled macrophage activation is found in the gut

IL-10 promoter microsatellites and SNPs

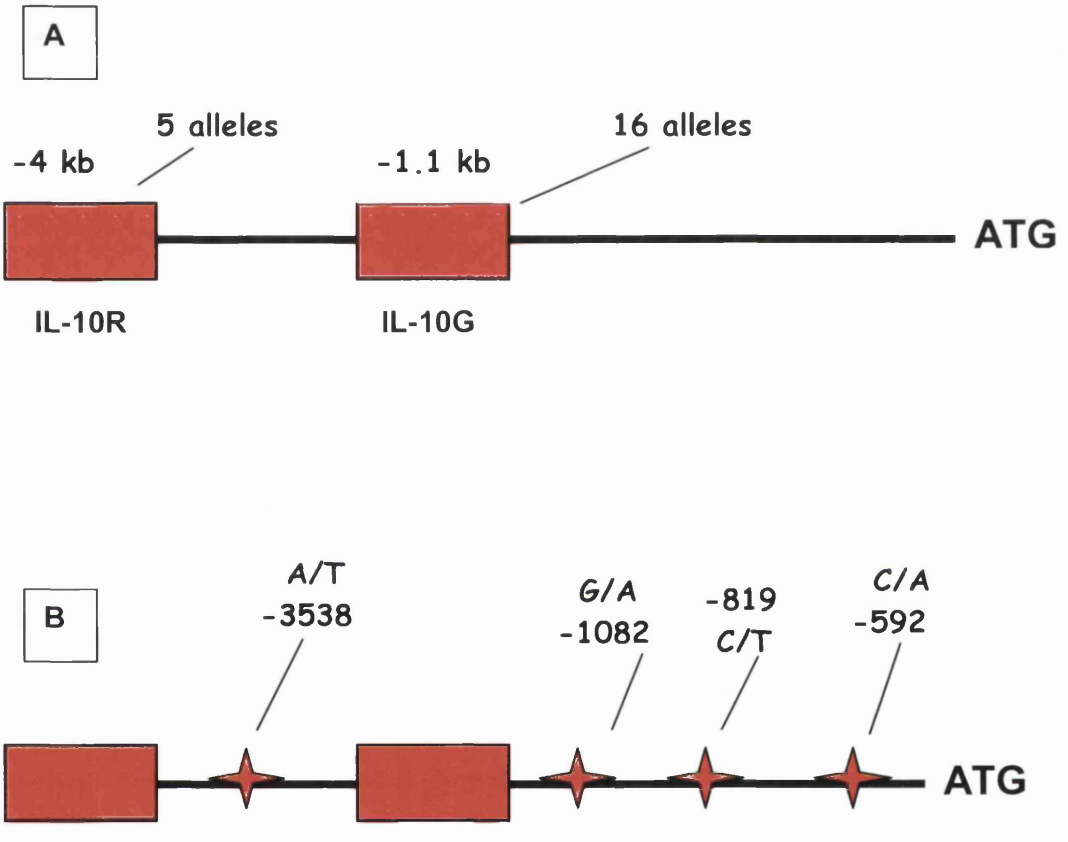


Fig 3.1.1 illustrates the promoter region of the IL10 gene.
Fig A shows the IL10 microsatellites IL10.R and IL10.G in the promoter region of the IL10 gene. IL10.R is 4kb 5' of the transcription initiation site and IL10.G 1.1 kb.
Fig B shows four of the single nucleotide polymorphisms (SNPs) identified in this region. To date, 10 SNPs have been identified but the SNPs illustrated here have been studied in some detail. Various studies have reported associations with different SNPs or haplotypes but the data has been inconsistent.

of these animals. This may lead to greater antigen presentation and massive exposure of lymphoid cells to luminal antigens. The disease is probably perpetuated by chronic over expression of cytokines. Mice kept under specific pathogen-free conditions develop only a local inflammation limited to the proximal colon suggesting the bowel inflammation may be a failure to control normal intestinal immune responses against enteric antigens and that IL10 is an essential immunoregulator in the intestinal tract. Repeated administration of IL10 into these mutants can transiently cure their disease (Kuhn *et al.*, 1993).

Variation in IL10 secretion is largely genetically determined (Westendorp *et al.*, 1997) and differences in secretion have been associated with various chronic inflammatory and infectious diseases. High levels of secretion are associated with a poor or fatal outcome in meningitis (Derkx *et al.*, 1995; Lehmann *et al.*, 1995; Westendorp *et al.*, 1997) and low levels are associated with susceptibility to severe malarial anaemia (Kurtzhals *et al.*, 1998).

3.1.3 Interleukin 10 gene and protein

IL10 is a homodimeric protein with subunits and is 160 amino acids in length. Murine IL10 shows 81% homology at the nucleotide level and 73% amino acid homology with human IL10 (de Waal Malefyt *et al.*, 1992; Kim *et al.*, 1992). Human IL10 is also closely related to the BCRF-1 gene (often called viral IL10 or vIL10) of the Epstein-Barr virus. Viral IL10 is more closely related at the protein level (84% homology) than murine IL10 (Vieira *et al.*, 1991).

The IL10 gene (appendix 5.5 for IL10 sequence) maps to the junction of 1q31-q32 (Eskdale *et al.*, 1997a). It contains four exons and exhibits

substantial polymorphism in the promoter region which appear to correlate with variation in transcription (Crawley *et al.*, 1999; Huizinga *et al.*, 2000) (fig 3.1.1). Two microsatellite polymorphisms, IL10.G and IL10.R, situated 1.1 and 4kb 5' of the transcription initiation site respectively are of particular interest (Eskdale and Gallagher, 1995; Eskdale *et al.*, 1996) since variation in IL10 secretion has been associated with particular haplotypes defined by these microsatellite markers (fig 3.1.1). Haplotypes containing the IL10.R2 allele are associated with a higher level of secretion than those including IL10.R3 (Eskdale *et al.*, 1998a). To date, eight single nucleotide polymorphisms (SNPs) have also been identified in the promoter region of this gene (D'Alfonso *et al.*, 2000; Eskdale *et al.*, 1999; Turner *et al.*, 1997), three of which have been studied in some detail: -1082 (G to A), -819 (C to T) and –592 (C to A) (Crawley *et al.*, 1999; Eskdale *et al.*, 1999; Turner *et al.*, 1997). Haplotype frequencies in 119 Caucasian controls were found to be 0.51 (GCC), 0.28 (ACC) and 0.21 (ATA) (Turner *et al.*, 1997). Studies investigating secretion of IL10 in relation to SNP polymorphisms have reported varying results. The –1082G allele has been associated with low levels of IL10 in one report (Huizinga *et al.*, 2000) but high levels in another (Turner *et al.*, 1997). A third study found decreased secretion in relation to the ATA haplotype (Crawley *et al.*, 1999). It is difficult to compare these results as the researchers employed different experimental protocols (Crawley *et al.*, 1999; Eskdale *et al.*, 1998a; Turner *et al.*, 1997) and the variation in IL10 production between different haplotypes is small in comparison to the large inter-individual differences (Huizinga *et al.*, 2000). Basal levels of IL10 secretion were higher when peripheral blood mononuclear culture (PBMC) cultures rather than whole blood cultures were used, probably reflecting stimulation of the cells when separated (Crawley *et al.*, 1999; de Groote and Zangerel, 1992). Cells can also be stimulated by lipopolysaccharide (LPS) stimulation (Crawley *et al.*, 1999) and concanavalin A (Con A) stimulation (Turner *et al.*, 1997).

Various genotypes of IL10.R, IL10.G and SNPs have been reported to show association with a variety of chronic inflammatory diseases. The ATA haplotype has been associated with both extended oligo-articular juvenile idiopathic arthritis (JIA) in Caucasians (Crawley *et al.*, 1999) and nephritis in a Southern Chinese population with systemic lupus erythematosis (SLE) (Mok *et al.*, 1998). The IL10.G microsatellite has also been associated with SLE (Eskdale *et al.*, 1997b) and more recently the IL10.R2 allele was found to be over-represented in rheumatoid arthritis whilst the IL10.R3 allele was under-represented (Eskdale *et al.*, 1998b).

3.1.4 Candidate gene studies

Association studies are generally used to study candidate genes. The simplest form of association study is the case control study (section 1.2.10.6). This has the advantage of relatively easy collection of 'cases' as other family members are not required. In genetic studies, where only DNA is needed from the control population, cells from normal blood donors can be used providing they are of the same ethnic background and sex matched as the cases. However, non-random mating in a population leading to population stratification or assortative mating can produce spurious results (section 1.2.10). Most polymorphisms studied are unlikely to be the real cause of disease susceptibility unless there is convincing evidence that the polymorphism is directly involved in influencing the production of a protein such as producing a stop codon and therefore a truncated protein. It is more likely that the polymorphism is a marker for other relevant mutations within the gene. Therefore, haplotyping is often undertaken to narrow down the region of interest (section 1.2.10.5). For instance, it is conceivable that only certain haplotypes containing IL10.R2 include the specific SNPs associated with increased IL10 secretion. The identification of allele transmission through families is required for haplotyping but as case control studies are

unable to provide this information other forms of family association studies are performed such as transmission disequilibrium (TDT) analysis (section 1.2.10.3).

Three association studies have been undertaken to investigate whether IL10, a biologically plausible candidate gene, is associated with susceptibility to or severity of RA. Two investigations were case control studies undertaken to demonstrate whether an association existed between RA susceptibility and IL10.R microsatellite alleles in two ethnically diverse populations (reprint in appendix). The third study used a family-based association method to undertake haplotyping of the IL10.R and IL10.G microsatellite markers in a cohort of UK Caucasians with RA. The haplotyping study was completed to determine whether specific haplotypes were associated with RA susceptibility.

Chapter 3 – section 2

3.2 Studies 1 and 2: Case Control Association

Rheumatoid arthritis susceptibility and interleukin 10: A study of two racially diverse populations.

3.2.1 Abstract

Introduction: IL10 is an immunoregulatory cytokine which may modulate disease expression in rheumatoid arthritis (RA). The IL10 gene is highly polymorphic with a number of single nucleotide polymorphisms in the promoter region and two microsatellite loci, IL10.R and IL10.G, 4kb and 1.1kb 5' of the transcription initiation site. It has been reported that allele 2 of the IL10.R microsatellite (IL10.R2) is associated with increase IL10 secretion and IL10.R3 with reduced secretion. Subsequently, over-representation of IL10.R2 and under-representation of IL10.R3 in three independent RA groups has been reported.

Aim: To determine whether there is an association between the IL10.R2 allele and RA in two ethnically distinct populations.

Methods: IL10.R genotypes were determined by semi-automated DNA sequencing technology in 186 UK Caucasians and 138 South Africans of Zulu or Sotho origin, fulfilling the 1987 ACR criteria for RA. The Caucasian patients had relatively severe disease and comprised 75 patients with RA vasculitis, 22 with Felty's Syndrome and 89 who had undergone a joint replacement (hip or knee) within 15 years of the onset of disease. Allele frequencies were compared with 296 Caucasians and/or 73 South Africans.

Results: The frequency of the IL10.R2 allele was significantly greater in the South Africans (RA and controls) than in the Caucasians (0.78 vs 0.66, $p=1\times10^{-6}$) while the frequency of IL10.R3 was less common (0.16 vs 0.3,

$p=1\times10^{-8}$). No differences were observed in either in IL10.R2 or IL10.R3 frequencies between patients and controls in either population.

Conclusions: No association between IL10.R alleles and RA was confirmed in this study. However, significant differences were demonstrated in the frequency of IL10.R2 and IL10.R3 between the two ethnic groups. The relatively high frequency of IL10.R2 in the South African population (0.78) would have reduced the power to detect an association with RA.

3.2.2 Introduction

IL10 is a biologically plausible candidate gene as described in the section above. As such, these studies were undertaken following the report by Eskadale and colleagues suggesting an association between the IL10.R2 allele and susceptibility to RA (Eskdale *et al.*, 1998b). The aim of these case control studies was to investigate further the association of IL10.R2 in UK Caucasian population with severe RA and to determine whether the association extended to another racially distinct population.

3.2.3 Materials and methods

3.2.3.1 Patients and controls

Two cohorts of racially distinct patients with RA and ethnically matched control groups were recruited from the United Kingdom and South Africa. The patients consisted of 186 Caucasian and 138 Black South African (SA) patients of Sotho or Zulu ethnicity, fulfilling the American College of

IL10.R microsatellite alleles assigned using the program GENOTYPER™

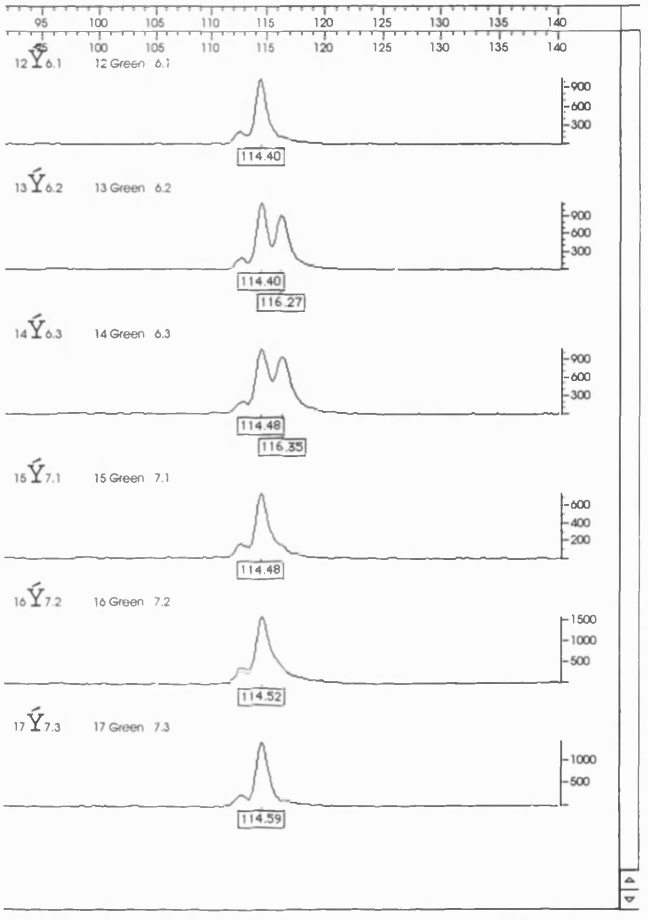


Fig 3.2.2 is an example of the output obtained using the program GENOTYPER™. Once the IL10.R microsatellite has been amplified by PCR, the products separated by electrophoresis and then sized by the program GENESCAN™, the genotypes are semiautomatically assigned using GENOTYPER™. The IL10.R2 product is 114 base pairs (bp) in length and IL10.R3 is 116 bp. Therefore two of the individuals shown in this example are IL10.R2 and IL10.R3 heterozygotes and the other four are IL10.R2 homozygotes.

Rheumatology (ACR) 1987 criteria for RA. The Sotho and Zulu groups have a similar genetic background (Spurdle and Jenkins, 1992). The SA patients with RA were recruited from the rheumatology outpatient clinic at the Chris Hani Baragwanath Hospital, Johannesberg, a tertiary referral centre. They were not chosen specifically for disease severity but all required disease modifying (DMARD) therapy. Dr M Tikly, their rheumatologist, recruited them for this study in 1997. The Caucasian patients were recruited from four centres within the UK: Oxford, Birmingham, Norfolk and Norwich and Glasgow by their rheumatologists', Professors P. Wordsworth, P Bacon, DGI Scott, R Sturrock and Dr. R Madhok. The UK Caucasian patients were recruited if they had relatively severe disease and comprised of 75 patients with RA vasculitis, 22 with Felty's Syndrome and 89 who had undergone a large joint replacement (hip or knee) within 15 years of the onset of disease. The mean age (± SD) of the SA patients was 49.5 (± 10.9) years, average disease duration was 7.6 (± 7.3) years and 76% were female. The mean age (± SD) of the Caucasian RA patients was 64 (± 11.6) years and 70 per cent were female. Mean disease duration was 18.7 (± 10.5) years and mean age of disease onset was 44.9 (± 14.5) years. The South African controls included 73 healthy, ethnically matched hospital workers or outpatients seen at the Baragwanath Hospital for minor trauma and unrelated to each other or to the cases. The UK Caucasian control group consisted of 210 healthy Caucasian blood donors and 86 unaffected Caucasian spouses of patients attending a skeletal dysplasia clinic.

3.2.3.2 IL10.R genotyping:

Once the patients had provided a blood sample the whole blood was frozen at -20°C prior to DNA extraction. Blood samples taken in South African were frozen there and brought to the UK on ice. Genomic DNA was then extracted from peripheral venous blood samples using a standard guanidine-hydrochloride extraction technique (section 5.2.1). Concentrated DNA was at stored -70°C following extraction and a small sample diluted with sterile water

to a concentration of 10ng/μL for the study. This dilute DNA was stored at 4°C for the duration (two weeks) of the study. The IL10.R microsatellite, was amplified by PCR (primers were 5' CCC TCC AAA ATC TAT TTG CAT A (upstream) and 5' CTC CGC CCA GTA AGT TTC ATC (downstream), the latter being tagged with a fluorescent dye (HEX)). Reactions were optimised and carried out in 96 well plates (Costar) in 10μL reactions consisting of 50ng DNA, 400nM each primer, 50μM each dNTP, 2.0mM $MgCl_2$ and 0.2 units DNA polymerase (Bioline, UK) in the manufacturers NH_4 buffer. The cycling conditions were 94°C 1 minute, annealing 60°C 1 minute, extension 72°C 1 minute, for 32 cycles on MJ thermal cyclers. PCR products were diluted with water and separated by electrophoresis using ABI 373 DNA sequencers (Applied Biosystems, Warrington, UK) and 6% denaturing polyacrylamide gels over 3 hours. Test gels were run to ensure the PCR dilutions were optimised prior to separating PCR products by electrophoresis. Products were sized using the program GENESCAN™ Version 2.1 (Applied Biosystems, Warrington, UK) and genotypes semiautomatically assigned using the program GENOTYPER™ Version 1.1 (Applied Biosystems, Warrington, UK) (figs 3.2.1 and 3.2.2). All genotypes were then verified manually. The program GAS (Version 2) (A. Young, unpublished) was used to convert the size data into discrete allele numbers.

3.2.3.3 HLA-DR typing:

Sequence specific PCR, using 35 primers, was used to differentiate between the different HLA-DR alleles and undertake DR4 and DR1 subtyping (Bunce *et al.*, 1995). Reactions were performed in 96 well plates (Costar) in 10μL reactions consisting of 50ng DNA, 400nM each primer, 50μM each dNTP, 2.0mM $MgCl_2$ and 0.2 units DNA polymerase (Bioline, UK) in the manufacturers NH_4 buffer. Cycling conditions were: 96°C (1min); 5 cycles (96°C (35sec), 70°C (45sec), 72°C (35sec)); 21 cycles (96°C (25sec), 65°C

(50sec), 72°C (40sec)); 6 cycles (96°C (35sec), 55°C (1 min), 72°C (1.5mins)); hold at 15°C. Total duration of PCR was 1 hour 30 minutes.

3.2.3.4 Statistical analysis:

Allele and genotype frequencies were calculated by direct counting. Since a common source of error in genotype assignment is the over-calling of homozygotes, Hardy-Weinberg equilibrium (section 1.2.5.2) was used to predict the likely frequencies of IL10.R2 and IL10.R3 homozygotes and these figures were compared with the observed frequencies. IL10.R allele frequency distribution was compared between the Caucasians and South Africans and then between patients and controls. Subgroup analysis included disease severity (RA vasculitis, Felty's Syndrome or an early large joint replacement), sex, and shared epitope status (homozygosity or heterozygosity). The significance of differences between groups was calculated from contingency tables by χ^2 analysis. Odds ratios with confidence intervals were calculated.

3.2.4 Results

The frequency of the IL10.R2 allele was significantly higher ($p=1\times10^{-6}$) in the South African population overall (0.78) compared to the UK Caucasians (0.66) while the frequency of IL10.R3 was correspondingly reduced (0.16 versus 0.30, $p=1\times10^{-8}$) (table 3.2.1). However, no differences were observed in IL10.R allele frequencies between patients and controls in either racial group (table 3.2.2). Eight-four per cent of the UK Caucasian patients and 58 per cent of the SA patients were positive for the shared epitope.

A variety of subgroups were defined from the UK Caucasian RA cohort to analyse any possible associations with IL10.R alleles. All alleles frequencies were very similar whether the groups were divided by sex, age of onset,

IL10.R gel

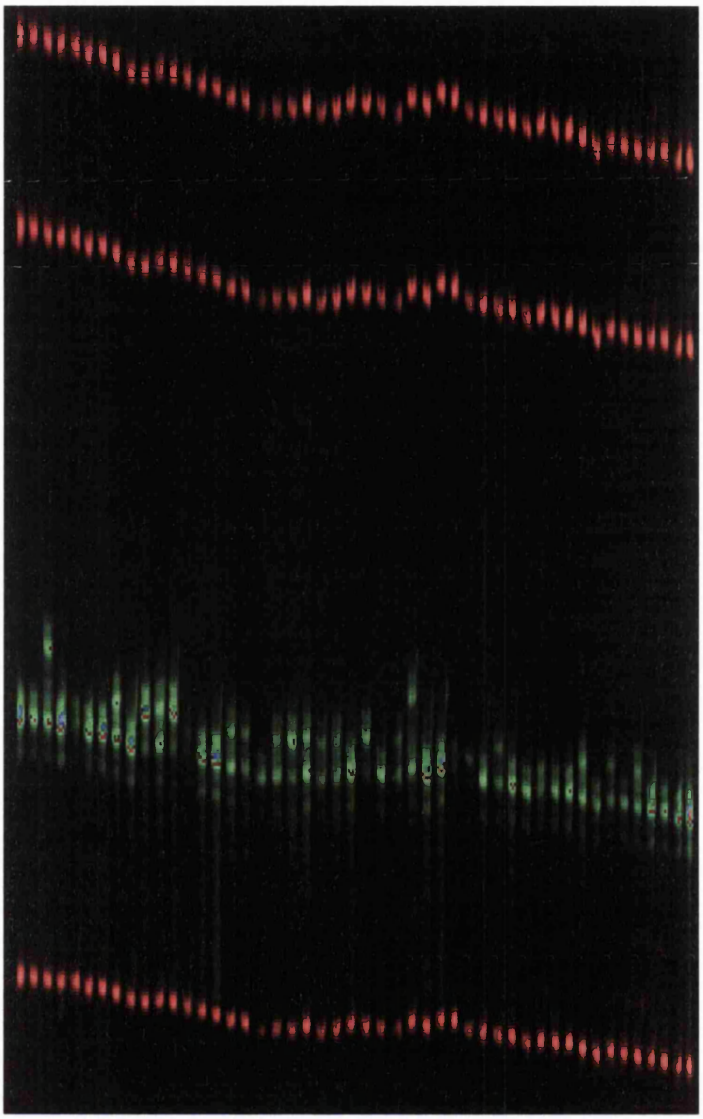


Fig 3.2.1 is an example of the output from an ABI 373 sequencer once the microsatellite marker IL10.R has been separated by electrophoresis over 3 hours using a 6% denaturing poylacrylamide gel. One of the IL10.R primers has a green fluorescent marker attached hence the amplified IL0.R microsatellite is shown as a green line on the gel. Products are sized using the program GENESA CAN" which compares the size of the green PCR product with that of the red fluorescent markers also seen in the figure. These are size markers and are loaded at the same time as the PCR products. They are spaced at regular intervals throughout the gel and are used as a form of ruler.

IL10.R allele frequencies in Caucasian and South African individuals

Comparison groups	IL10.R alleles	Alleles (Caucasian patients and controls)		Alleles (South African patients and controls)		P
		No.	%	No.	%	
Caucasian versus South Africans	IL10.R1	4	0.4	11	2.5	ND
	IL10.R2	640	66	350	79	1×10^{-6}
	IL10.R3	287	30	69	16	1.4×10^{-8}
	IL10.R4	32	3.3	11	2.6	0.4
	IL10.R5	1	0.1	1	0.3	0.6

Table 3.2.1 compares the allele frequencies of the various IL10.R alleles in the Caucasian and South African populations. It can be seen that the IL10.R2 allele occurs more commonly in the South African population than the Caucasian population (0.79 versus 0.66) and the IL10.R3 allele less commonly (0.16 versus 0.3).

IL10.R allele frequencies in Caucasian and South African patient and control groups.

Comparison groups	IL10.R alleles	Patients		Controls		P	OR (95%ci)
		No.	%	No.	%		
Caucasian patients versus controls	IL10.R1	3	0.8	1	0.2	ND	
	IL10.R2	253	68	387	65	0.4	1.3(0.8-1.5)
	IL10.R3	103	28	184	31	0.2	0.85(0.6-1.4)
	IL10.R4	13	2.7	19	3.2	0.8	
	IL10.R5	0	0	1	0.2	ND	
South African patients versus controls	IL10.R1	9	3.2	2	1.4	ND	
	IL10.R2	223	78	127	77	0.4	0.8(0.5-1.4)
	IL10.R3	47	17	22	15	0.5	1.25(0.7-2.2)
	IL10.R4	7	2.5	4	2.7	ND	
	IL10.R5	0	0	1	0.7	ND	

Table 3.2.2 compares the allele frequencies of the various IL10.R alleles in the Caucasian and South African patients versus the relevant control populations. No differences in allele frequencies were found between the South African patients and South African controls or between the Caucasian patients and controls.

shared epitope status, extra-articular disease (RA vasculitis and Felty's Syndrome) or early large joint replacement (table 3.2.3). The observed frequencies of IL10.R2 and IL10.R3 homozygous genotypes compared well with the predicted frequency of homozygotes by Hardy-Weinberg equilibrium suggesting alleles were being appropriately assigned.

3.2.5 Discussion

No association between IL10 haplotypes and RA was apparent in the current study, in contrast to previous reports of an increase in IL10.R2 and a reduction in IL10.R3 alleles (Eskdale *et al.*, 1998b). There are several possible explanations for this. Firstly, the previously reported association with the IL10.R2 allele may have been spurious. Secondly, it may have been relevant that the IL10.R2 allele frequency in UK Caucasian controls in the Oxford study were significantly higher (0.65 versus 0.59, p=0.03) than that reported by Eskdale et al and the IL10.R3 frequency significantly lower (0.31 versus 0.38, p=0.02). No difference in allele frequencies of IL10.R2 or IL10.R3 was apparent when the UK Caucasian patient groups from the two studies were compared (table 3.2.4). This may suggest that IL10 contributes a weak genetic effect but the relatively high frequency of IL10.R2 in the general population made the effect more difficult to detect reliably. Since the frequency of IL10.R2 was even higher in South African Sotho and Zulu populations, the power to detect association with RA would be further reduced. Previously reported estimates of relative risk for IL10.R in RA range between 1.5 and 2.5 (Eskdale *et al.*, 1998b). The study in UK Caucasians had 80 per cent power to exclude an association between IL10 and RA with an odds ratio of $\geq$ 1.8. However, in the South African study the higher frequency of the IL10.R2 allele (SA control frequency 77% versus 65% in Caucasians) meant that the power to detect an effect with an odds ratio of $\geq$ 1.8 was only 36 per cent. In contrast the study had 80 per cent power to

IL10.R allele frequencies in Caucasian patient subgroups versus Caucasian control populations

Comparison groups	IL-10R alleles	Case alleles		Control alleles		P
		No.	%	No.	%	
Early large joint replacement	IL-10R1	0	(0)	1	(0.2)	ND
	IL-10R2	123	(69)	387	(65)	0.4
	IL-10R3	49	(28)	184	(31)	0.4
	IL-10R4	6	(3.3)	19	(3.2)	0.9
	IL-10R5	0	(0)	1	(0.2)	ND
RA vasculitis	IL-10R1	3	(2)	1	(0.2)	ND
	IL-10R2	100	(67)	387	(65)	0.07
	IL-10R3	42	(28)	184	(31)	0.46
	IL-10R4	5	(3.3)	19	(3.2)	0.9
	IL-10R5	0	(0)	1	(0.2)	ND
Felty's Syndrome	IL-10R1	0	(0)	1	(0.2)	ND
	IL-10R2	29	(66)	387	(65)	0.9
	IL-10R3	13	(29)	184	(31)	0.8
	IL-10R4	2	(5)	19	(3.2)	ND
	IL-10R5	0	(0)	1	(0.2)	ND

Table 3.2.3 compares the IL10.R allele frequencies in the Caucasian patient subgroups with the Caucasian control population. Allele frequencies were compared using χ^2 analysis. No differences in allele frequencies were found.

exclude a putative genetic effect with an odds ratio of ≥ 3. Thirdly, genetic heterogeneity may be operating. The UK patients in this study were specifically selected for having more severe forms of RA and were only included if they had undergone an early large joint replacement or fulfilled the criteria for rheumatoid vasculitis or Felty's Syndrome. This was in contrast to the previous report where the main recruitment criteria was RA fulfilling the 1987 ACR criteria (Eskdale *et al.*, 1998b). It is therefore likely that the disease severity of the Caucasian patients included in the two studies was different and it is conceivable that the IL10.R2 allele is not as strongly associated with severe forms of RA as it is with milder variants.

Inter-ethnic differences in IL10 allele frequencies were not unexpected as similar differences have been described previously for the TNF locus (Fei *et al.*, 2002; Gallagher *et al.*, 1997; Rudwaleit *et al.*, 1996). Although in the case of TNF, linkage disequilibrium within the MHC may contribute to this variation (Rudwaleit *et al.*, 1996).

A number of studies in normal individuals have demonstrated associations between IL10 secretion and microsatellite or SNP polymorphisms characterising distinct IL10 haplotypes (Eskdale *et al.*, 1998a; Turner *et al.*, 1997). Other studies have suggested that some of these haplotypes are associated with inflammatory diseases (Crawley *et al.*, 1999; Eskdale *et al.*, 1997b; Mozzato-Chamay *et al.*, 2000). However, there is no really convincing evidence that the IL10.R2 or IL10.R3 polymorphisms are directly involved in influencing IL10 production or disease susceptibility. They may be markers for other relevant mutations within the gene. It is possible that only certain haplotypes containing IL10.R2 include the specific SNPs associated with increased IL10 secretion and susceptibility to RA. Equally, only certain haplotypes containing IL10.R3 may truly be under-represented and associated with reduced IL10 secretion. Discrimination between the various IL10.R2 or IL10.R3 extended haplotypes is not possible in a case control

Comparing IL10.R2 and IL10.R3 allele frequencies from published studies

Study	Patient group	IL10.R2 % RA	IL10.R2 % control	IL10.R 3 % RA	IL10.R3 % control
Rheumatology 2003 (MacKay et al)	Caucasians with severe RA	68	65	28	31
Rheumatology 2003 (MacKay et al)	Black South Africans (disease severity – not documented)	78	77	17	15
Lancet 1998 (Eskdale et al)	Oxford Caucasians (disease severity – not documented)	70	61	25	36
Lancet 1998 (Eskdale et al)	Glasgow Caucasians (disease severity – not documented)	69	56	29	40
Lancet 1998 (Eskdale et al)	African Americans (disease severity – not documented)	87	72	11	24

Table 3.2.4 summarises the published IL10.R2 and IL10.R3 allele frequencies in different patient groups (3 Caucasian, 1 African American and 1 South African) and control populations. It can be seen that the Caucasian control population used in this study (row 1 of the table) has a higher IL10.R2 allele frequency than other Caucasian control populations.

study. As current evidence regarding a possible association of IL10 with RA is inconclusive it would be appropriate to conduct a within-family association study to define the effects of specific haplotypes and avoid any possible spurious effects that could occur as a result of unrecognised population stratification.

In conclusion, these studies did not confirm an association between IL10.R2 and susceptibility to RA. However, as the data was inconclusive further studies investigating the likelihood that IL10 is associated with susceptibility to or severity of RA would be justified. To achieve adequate statistical power studies involving a larger cohort of patients would be required. Extended haplotyping of the IL10 promoter region should help to define any disease causing haplotypes and so improve the chances of identifying an association.

Chapter 3 – section 3

3.3 Study 3: a family-based association study.

Interleukin 10 haplotypes are associated with rheumatoid arthritis.

3.3.1 Abstract

Introduction: Interleukin 10 (IL10), an immunoregulatory cytokine, may modulate disease expression in rheumatoid arthritis (RA). The IL10 gene is highly polymorphic with a number of single nucleotide polymorphisms and two microsatellite loci, IL10.R and IL10.G, in the promoter region. A recent report has suggested that allele 2 of the IL10.R microsatellite (IL10.R2) is associated with increased IL10 secretion and IL10.R3 with reduced secretion. Another report has found over-representation of IL10.R2 in three independent RA cohorts and under-representation of IL10.R3.

Aim: To determine by within family association analysis whether particular alleles or haplotypes of IL10.R and/or IL10.G are associated with RA.

Methods: The IL10.R and IL10.G microsatellites were used to genotype 163 individuals, fulfilling the 1987 ACR criteria for rheumatoid arthritis, and their first-degree relatives. Single marker and haplotypic association studies were performed by transmission disequilibrium testing (TDT) using the software package TRANSMIT.

Results: The IL10.R1 allele was transmitted to affected individuals more frequently than expected ($p<0.05$) and the IL10.R3 allele was transmitted less frequently than expected ($p<0.05$). This effect was particularly pronounced with the IL10.R3 / IL10.G10 haplotype ($p<0.005$).

Conclusions: This study is consistent with previous observations that the IL10.R3 allele is under-represented in RA. However, the still stronger negative association found with the haplotype IL10.G10 / IL10.R3 suggests it is not IL10.R itself but another polymorphism on this particular haplotype that may be primarily involved with RA.

3.3.2 Introduction

Results from case control studies investigating the association between RA susceptibility and IL10 microsatellite alleles (see above) have been conflicting. There is no really convincing evidence that IL10 polymorphisms are directly involved in influencing IL10 production or disease susceptibility. The IL10 polymorphisms could be markers for other relevant mutations within the gene causing true disease susceptibility or severity. Distinguishing between a primary effect arising from a genetic variant or an effect arising from linkage disequilibrium can be difficult using a case-control study. Also, type 1 errors (false positive) can occur as a result of unidentified population stratification (section 1.2.10). Family-based association studies avoid population stratification by using family members as internal controls (i.e. the non-transmitted allele is the 'control') and also allow for the analysis of haplotype transmission. As haplotypes mark recognisable chromosomal segments that are transmitted en-block through a pedigree this may increase the likelihood of identifying a true association. Hence a third study was designed to investigate a possible association between the IL10 locus and susceptibility to rheumatoid arthritis by using intra-familial methods of association analysis including transmission disequilibrium testing (TDT analysis).

Information regarding the background work relating to IL10 and its candidacy for association with RA susceptibility has been discussed earlier in this chapter.

3.3.3 Materials and methods

3.3.3.1 Patients and controls

One hundred and sixty-three simplex Caucasian RA families were studied. All were recruited from United Kingdom; 55 from the Oxfordshire area and the remaining 108 from the Arthritis Research Campaign (ARC) United Kingdom National Repository of multicase RA families (ARC multicase families). All affected individuals fulfilled the American College of Rheumatology 1987 criteria for RA. Families recruited in the Oxfordshire area were recruited if blood was available from the proband and both parents (i.e. TDT families). Local ethics committee approval was obtained and all patients and probands gave fully informed consent. The families were recruited in 1997 by a research nurse based at the Wellcome Trust Centre for Human genetics (WTCHG) in Oxford. DNA was available from the proband and both parents in 108 (66%) of families (from the Oxfordshire TDT families and 38 ARC multicase families) but only one parent and the proband were available for the remaining 55 (34%) families. Additional siblings were recruited where possible (14/55 families) to facilitate the assignation of parental genotypes.

3.3.3.2 Probands

The mean age (+/- SD) of the probands was 38 (+/- 10) years. Eighty-two percent of the probands were female. Forty-four percent of the probands included in the study had a first-degree relative (62 parents and 11 siblings) with rheumatoid arthritis. Twenty-one (15%) probands had an affected father and 41 (25%) an affected mother. There was no difference in the frequency

IL10.G and IL10.R gel

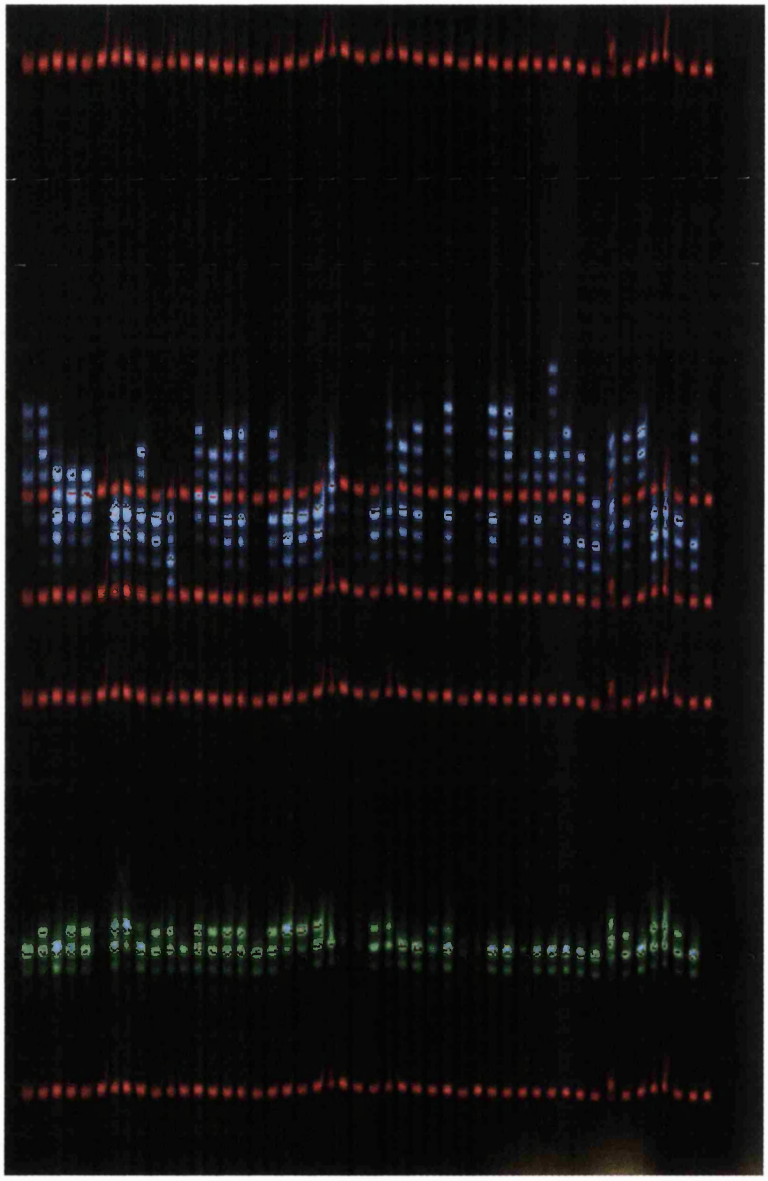


Fig 3.3.1 is an example of the output from an ABI 373 sequencer once the microsatellite markers IL10.R and IL10.G have been separated by electrophoresis using a 6% denaturing poylacrylamide gel over 3 hours. The IL10.G PCR product has been tagged with a blue fluorescent label, IL10.R is green and the size markers are red (as in fig 3.2.1).

of having an affected parent between male and female probands (34% and 38% respectively, p = 0.83). An equal number of male probands had an affected father (4 probands, 14%) or an affected mother (4 probands). However, more female probands had an affected mother (34 probands, 25%) than an affected father (17 probands, 13%) (p=0.008).

3.3.3.3 IL10.G and IL10.R genotyping:

Genomic DNA was extracted from peripheral venous blood samples using a standard guanidine-hydrochloride extraction technique (appendix 5.2.1). Concentrated DNA was at stored -70°C following extraction and a small sample diluted with sterile water to a concentration of 10ng/μL for the study. This dilute DNA was stored at 4°C for the duration (two weeks) of the study. The IL10.R and IL10.G microsatellites were amplified by PCR. Primers for IL10.R were 5' CCC TCC AAA ATC TAT TTG CAT A (upstream) and 5' CTC CGC CCA GTA AGT TTC ATC (downstream) and IL10.G, 5' GAA GAA GTC CTG ATG TCA CT (upstream) and 5' GCC TTA GTA GTG TTG TCT TGG AT (downstream). The IL10.R primers used were the same as those used for the case-control studies (above) and as in the study by Eskdale et al (Eskdale *et al.*, 1998b). The IL10.G primers were designed for the study (K.R MacKay) and both sets of primers were supplied by MWG-Biotech, UK. Both reverse primers were tagged with a fluorescent dye (HEX or FAM). Reactions were optimised and performed in 96 well plates (Costar) in 10μL reactions consisting of 50ng DNA, 400nM each primer, 50μM each dNTP, 2.0mM $MgCl_2$ (4.0mM $MgCl_2$ for IL10.G) and 0.2 units DNA polymerase (Bioline, UK) in the manufacturers NH_4 buffer. The cycling conditions were 94°C 1 minute, annealing 60°C 1 minute, extension 72°C 1 minute, for 32 cycles. PCR products were diluted with water, pooled and separated by electrophoresis using an ABI 373 semi-automated sequencer (Applied Biosystems, Warrington, UK) and 6% denaturing polyacrylamide gels over 3 hours (fig 3.3.1). Test gels were run to ensure the PCR dilutions were optimised prior to separating PCR products by electrophoresis. Products were sized using

IL10.G microsatellite alleles assigned using the program GENOTYPER™

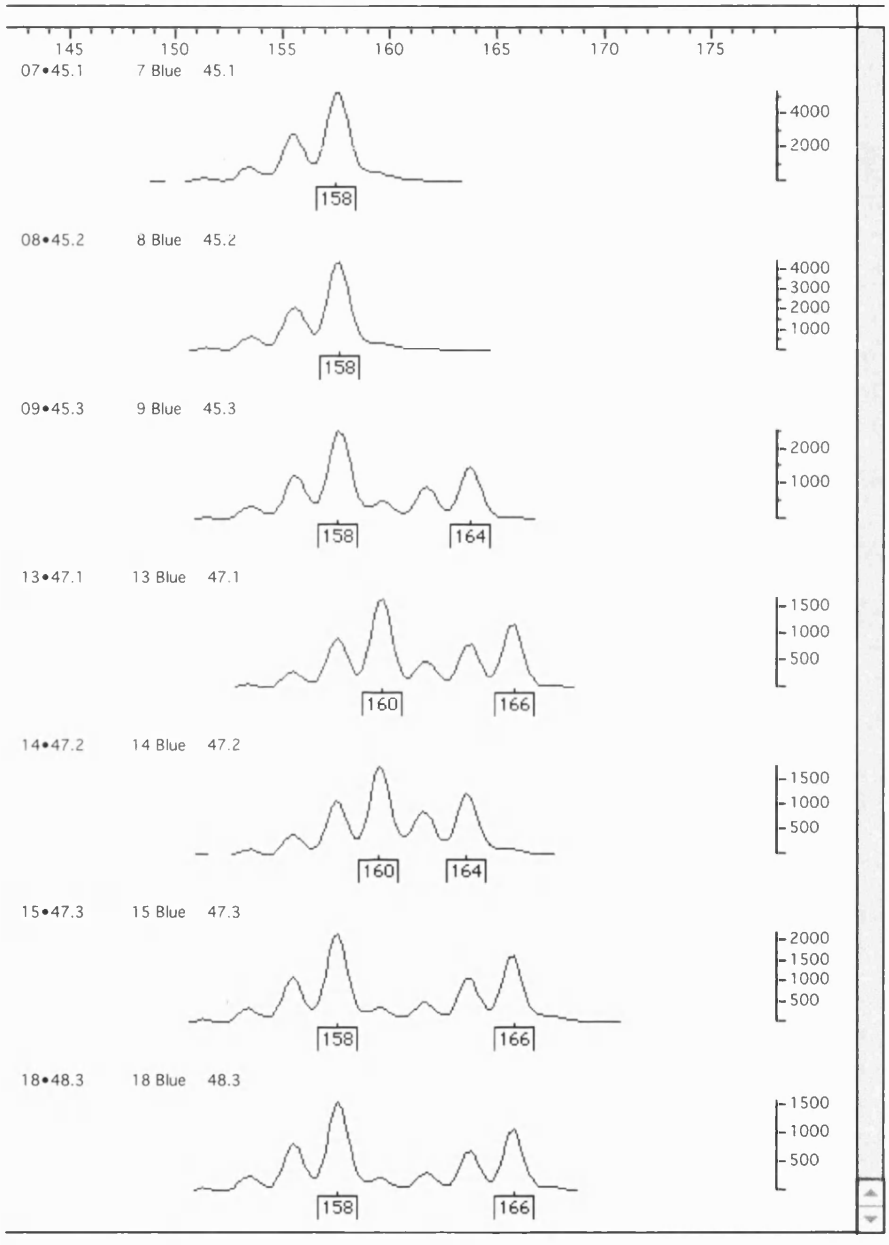


Fig 3.3.2 is an example of the output obtained using the program GENOTYPER™. The genotypes are semi-automatically assigned using GENOTYPER™. The PCR product sizes of the IL10.G alleles range from 142 – 172 base pairs in length.

the program GENESCAN™ Version 2.1 (Applied Biosystems) and genotypes semiautomatically assigned using the program GENOTYPER™ Version 2.1 (Applied Biosystems) (fig 3.3.2). All genotypes were then verified manually. The program GAS (Version 2) (A. Young, unpublished) was used to convert the size data into discrete allele numbers.

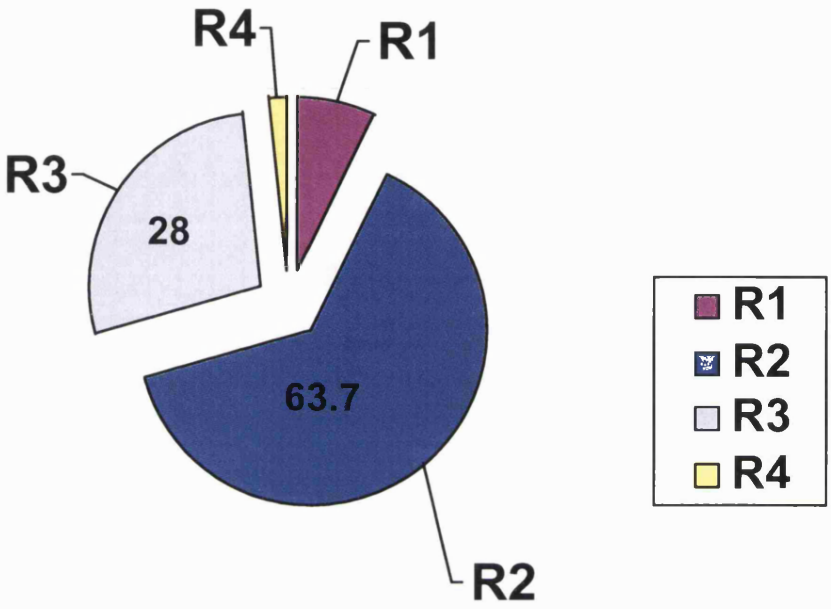
3.3.3.4 HLA-DR typing:

Sequence specific PCR, using 35 primers, was used to differentiate between the different HLA-DR alleles and undertake DR4 and DR1 subtyping (Bunce *et al.*, 1995; Olerup and Zetterquist, 1992). Reactions were performed in 96 well plates (Costar) in 10μL reactions consisting of 50ng DNA, 400nM each primer, 50μM each dNTP, 2.0mM $MgCl_2$ and 0.2 units DNA polymerase (Bioline, UK) in the manufacturers NH_4 buffer. Cycling conditions were: 96°C (1min); 5 cycles (96°C (35sec), 70°C (45sec), 72°C (35sec)); 21 cycles (96°C (25sec), 65°C (50sec), 72°C (40sec)); 6 cycles (96°C (35sec), 55°C (1 min), 72°C (1.5mins)); hold at 15°C. Total duration of PCR was 1 hour 30 minutes.

3.3.3.5 Statistical analysis:

Allele and genotype frequencies were calculated by direct counting. A common source of error in genotype assignment is the over-calling of homozygotes. Hardy-Weinberg equilibrium was used to predict the likely homozygote frequencies of the commonest IL10.R and IL10.G alleles (section 1.2.5.2). These figures were then compared with the observed frequencies. Single marker and haplotypic association studies were performed by transmission disequilibrium testing (TDT analysis) using TRANSMIT (section 1.2.10.3).

IL-10R allele frequencies



Alleles	Frequency
IL10.R1	0.072
IL10.R2	0.637
IL10.R3	0.28
IL10.R4	0.016
IL10.R5	0

Fig 3.3.3 illustrates the IL10.R allele frequencies identified in this Caucasian population. The most frequent allele was IL10.R2 with a frequency of 0.63. The second most frequent allele was IL10.R3 at 0.28.

3.3.4 Results

IL10.R and IL10.G genotypes and hapolotypes

IL10.R and IL10.G allele frequencies for the 458 individuals studied are shown in figures 3.3.3 and 3.3.4. The individual allele frequencies were similar to those previously reported, although IL10.R2 occurred more frequently (63%) and IL10.G9 less frequently (37%) than in some other studies. The existence of 12 alleles for IL10.G ranging from 14 CA repeats to 29 CA repeats was confirmed. However no IL10.G allele with a length of 15, 16, 17 or 18 CA repeats was found (fig 3.3.4 and table 3.3.1). This accounted for the apparent shortfall from the theoretical number of 16 alleles. If the missing alleles exist their frequency in Caucasians is less than 0.001.

Association testing for IL10.R and IL10.G as individual markers was possible in 157 families (IL10.G) and 162 families (IL10.R). TDT analysis using the software program TRANSMIT (section 1.2.10.3) identified an association for IL10.R (global χ^2, (3 degrees of freedom (df) = 8.7; $p < 0.05$). IL10.R1 alleles were transmitted more frequently than expected ($\chi^2 = 4.3$(1df); $p < 0.025$) and IL10.R3 alleles less frequently than expected ($\chi^2 = 4.1$(1df); $p < 0.025$) (table 3.3.2). IL10.G was not found to be associated with RA (global $\chi^2 = 9.6$(11df), NS).

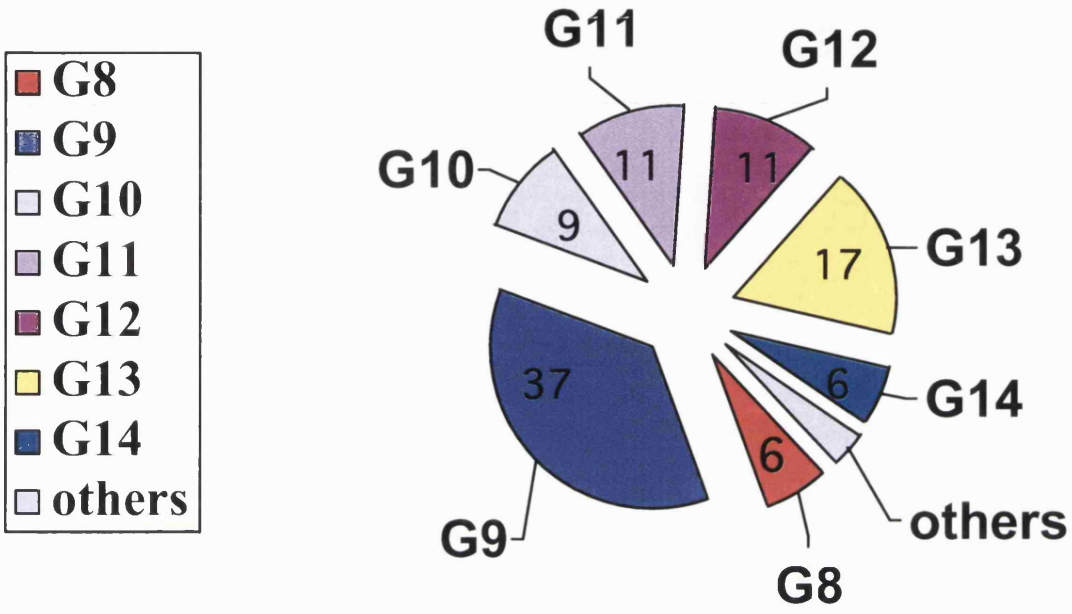
The frequencies of IL10.R/IL10.G haplotypes are shown in figure 3.3.5. The commonest haplotype was IL10.R3/IL10.G9 at a frequency of 18.7%. TDT haplotypic analysis using TRANSMIT revealed a strong association with the haplotype IL10.R3/IL10.G10 (χ^2=8.835 (1df), $p<0.005$) and a slightly weaker association with IL10.R1/IL10.G12 (χ^2= 5.1 (1df), $p<0.025$) (table 3.3.3). The haplotypic association for IL10.R3 / IL10.G10 was stronger than when any of the IL10.R alleles were analysed individually (table 3.3.2).

Differing PCR product sizes depending upon the amplifying primers

IL10.G allele number	Number of CA repeats	PCR product size (using original IL10.G primers)	PCR product size (using newly designed IL10.G primers)
IL10.G1	14	124	142
IL10.G2	15	126	144
IL10.G3	16	128	146
IL10.G4	17	130	148
IL10.G5	18	132	150
IL10.G6	19	134	152
IL10.G7	20	136	154
IL10.G8	21	138	156
IL10.G9	22	140	158
IL10.G10	23	142	160
IL10.G11	24	144	162
IL10.G12	25	146	164
IL10.G13	26	148	166
IL10.G14	27	150	168
IL10.G15	28	152	170
IL10.G16	29	154	172

Fig 3.3.1 lists two differing sets of IL10.G PCR product size. The size of the products depended upon which set of primers were used to amplify the IL10.G microsatellite. The Initial primers used to amplify IL10.G were those published by Eskdale et al [Eskdale, 1997 #532] but the primers were re-designed because of frequent primer failure. The re-designed primers amplified a larger segment of DNA, hence the product size was slightly larger.

IL-10G allele frequencies



Allele	Frequency	Allele	Frequency
IL10.G1	0.002	IL10.G9	0.366
IL10.G2	0	IL10.G10	0.095
IL10.G3	0	IL10.G11	0.108
IL10.G4	0	IL10.G12	0.108
IL10.G5	0	IL10.G13	0.169
IL10.G6	0.001	IL10.G14	0.058
IL10.G7	0.022	IL10.G15	0.005
IL10.G8	0.063	IL10.G16	0.005

Fig 3.3.4 illustrates the frequency of IL10.G alleles in this Caucasian population. The most frequent allele was IL10.G9 with a frequency of 0.366. The second most frequent allele was IL10.G13 at 0.169. The table lists the frequency of all the IL10.G alleles in this population.

3.3.5 Discussion

IL10.R was found to be associated with RA whereas IL10.G was not. This is in keeping with the results of the case control study reported by Eskdale and colleagues (Eskdale *et al.*, 1998b) but it is in contrast to the results of the two case control studies discussed above. The IL10.R3 allele was under-represented in this rheumatoid population and excess transmission of the less common IL10.R1 allele was identified. The under-representation of IL10.R3 corresponds to a previous study (Eskdale *et al.*, 1998b) but the over-representation of IL10.R1 has not been reported previously. Finally, the over-representation of IL10.R2 that was noted in a previous study was not confirmed (Eskdale *et al.*, 1998b). Haplotype analysis revealed a stronger negative association with the haplotype IL10.R3 / IL10.G10. This suggests that the locus IL10.R is not the primary association but that another polymorphism on the IL10.R3 / IL10.G10 haplotype may be involved.

The results of IL10 microsatellite association studies may vary because of differences in disease severity between the patient cohorts. A number of groups have reported associations between IL10 and disease severity rather than susceptibility [Mok, 1998 #327; Huizinga, 2000 #634]. In addition to any anti-inflammatory properties, IL10 plays an important role in B cell proliferation and differentiation (Rousset *et al.*, 1992). High levels of IL10 have been found in rheumatoid arthritis and may actively contribute to disease progression by elevating autoimmune activity (Perez *et al.*, 1981). IL10.R3 alleles have been associated with lower levels of IL10 secretion in vitro (Eskdale *et al.*, 1998a) and it may be of importance that IL10.R3 alleles have been found to be under-represented in this family association study and an earlier case-control study (Eskdale *et al.*, 1998b). Although, the cohort in this study was not specifically chosen for disease severity, the individuals developed RA at a relatively young age and many of them had a family history of RA. Both early age of onset and family history are risk factors for

TDT analysis of IL10.R alleles

Allele	Observed	Expected	χ^2 (1df)	P value
R1	28	22	4.3	<0.025
R2	222	217	0.6	NS
R3	86	98	4.1	<0.025
R4	6	5	0.6	NS

1df = 1 degree of freedom

Table 3.3.2 summarises the results of transmission disequilibrium testing (TDT) of IL10.R alleles. It can be seen that the IL10.R1 allele has a low population frequency but was inherited more frequently than expected. The IL10.R3 allele (second most frequent allele) was inherited less frequently than expected. The IL10.R1 results need to be treated with caution because the allele is rare and therefore the difference in allele frequency between 'observed' and 'expected' is small and may be prone to bias.

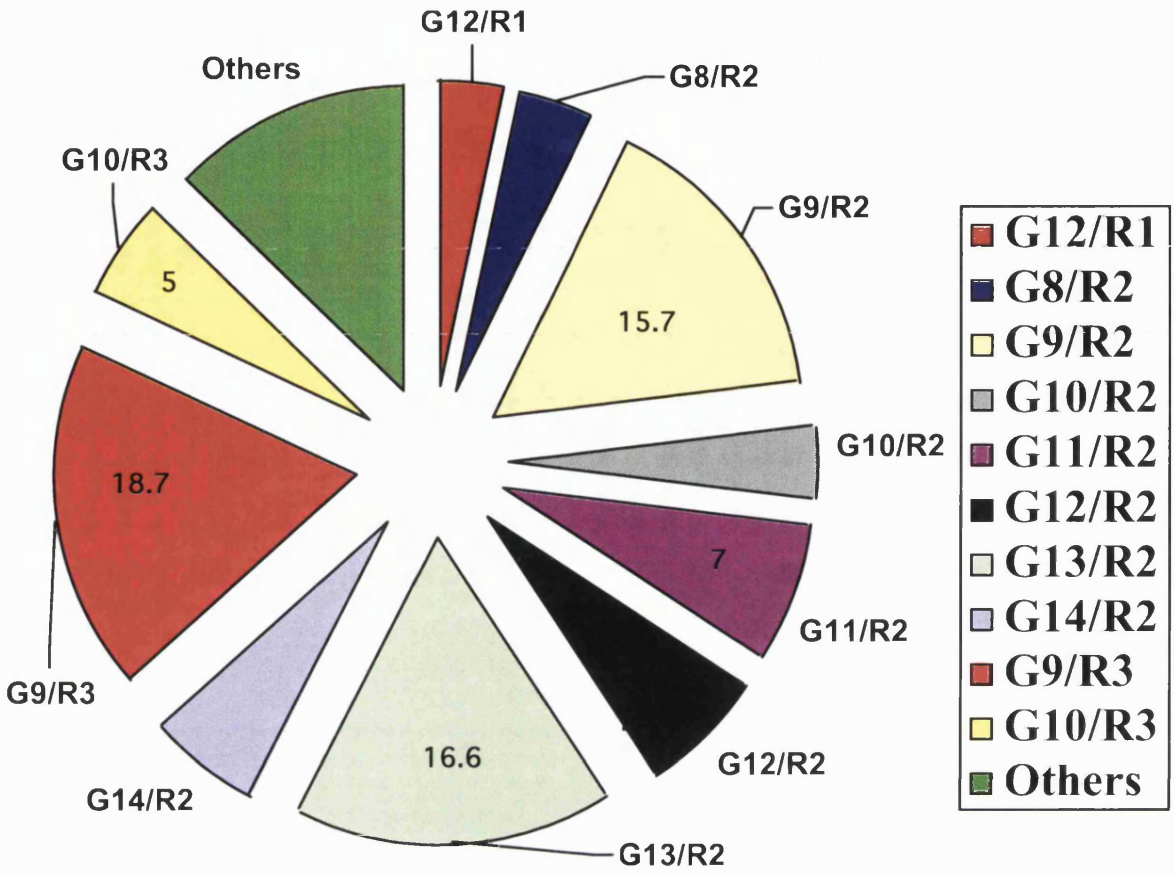
more severe disease (Deighton *et al.*, 1994; Grennan and Sanders, 1988; Lynn *et al.*, 1995).

An alternative explanation for these differing results could be the high IL10.R2 allele frequency in the general population as this may make the detection of a weak genetic effect unreliable. Small differences in the control IL10.R2 allele frequency could therefore alter results. Table 3.2.4 lists the IL10.R2 allele frequencies of the patient and control groups from the various IL10 microsatellite studies. The IL10.R2 allele frequencies of the Caucasian control populations do vary but the only statistically significant difference is between the controls used in the case-control study described above and those used by Eskdale and colleagues (0.65 versus 0.59, p=0.03) (Eskdale *et al.*, 1998b). The 'control' IL10.R2 allele frequencies in the TDT study do not differ statistically from those in either of the two case-control studies.

Although IL10.R has now been associated with RA in two independent studies, different IL10.R alleles have been identified as being associated with the disease in each report. Unidentified population stratification in a case control study could lead to a type 1 or type 2 error and it is possible that the reported association with the IL10.R2 allele is spurious. A family based association study avoids population stratification by using the un-transmitted parental allele as the 'control' hence avoiding one potential source of error. As both studies identified an under-representation of the IL10.R3 allele, on present evidence this seems to be the most likely true association.

The use of IL10 promoter polymorphisms to investigate the relationship of the IL10 gene with RA is based upon two assumptions. Firstly, that these polymorphisms are associated with differences in IL10 production. Secondly, that differences in IL10 secretion are associated with RA susceptibility or severity. However the relevant stimulus for the production of IL10 is unknown and the polymorphisms studied may be in linkage disequilibrium with other

IL10.R & G haplotype frequencies



IL10.R & G haplotype	Frequency	IL10.R & G haplotype	Frequency
IL10.G8/R1	0.016	IL10.G12/R2	0.066
IL10.G10/R1	0.002	IL10.G13/R2	0.166
IL10.G11/R1	0.009	IL10.G14/R2	0.057
IL10.G12/R1	0.032	IL10.G15/R2	0.003
IL10.G13/R1	0.012	IL10.G16/R2	0.003
IL10.G5/R2	0.002	IL10.G8/R3	0.009
IL10.G6/R2	0.002	IL10.G9/R3	0.187
IL10.G7/R2	0.023	IL10.G10/R3	0.053
IL10.G8/R2	0.041	IL10.G11/R3	0.022
IL10.G9/R2	0.157	IL10.G12/R3	0.010
IL10.G10/R2	0.039	IL10.G9/R4	0.010
IL10.G11/R2	0.074	IL10.G10/R4	0.002

Fig 3.3.5 illustrates the frequency of the commonest IL10.G and R haplotypes in this Caucasian population. The most frequent haplotype was IL10.G9/R3 with a frequency of 0.187. The table lists the frequency of all the IL10.G and R haplotypes within this population.

relevant mutations within the gene. Hence, the polymorphisms studied to date may not be truly associated with RA and defining a common DNA sequence transmitted en-block that is inherited by those with the disease would improve the chances of identifying a true disease associated polymorphism. Extended haplotyping in a family based association study can narrow down the DNA sequence of interest by identifying strong associations between specific haplotypes and RA. Therefore, the stronger negative association with the haplotype IL10.R3 / IL10.G10 suggests that the locus IL10.R is not the primary association and that another polymorphism on the IL10.R3 / IL10.G10 haplotype may be involved.

In conclusion, IL10 haplotypes are associated with RA. The IL10.R3 allele is under-represented and the IL10.R1 allele over-represented in individuals with RA. The haplotype IL10.R3 / IL10.G10 is more strongly associated suggesting that a polymorphism on this particular haplotype may be primarily involved with disease. Extended haplotyping, using both microsatellites and SNPs in a further independent family dataset, would be the next step to confirm the association and narrow down the DNA sequence of interest.

TDT analysis of IL10 G and R haplotypes

Haplotype	Observed	Expected	χ^2 (1df)	P value
IL10.G12/R1	14	10	5.7	<0.025
IL10.G8/R2	13	13	0	NS
IL10.G9/R2	52	52	0	NS
IL10.G10/R2	15	12	1.4	NS
IL10.G11/R2	29	24	2.6	NS
IL10.G12/R2	21	22	0	NS
IL10.G13/R2	52	55	0.5	NS
IL10.G14/R2	20	19	0	NS
IL10.G9/R3	61	61	0	NS
IL10.G10/R3	9	17	8.8	<0.005

1df = 1 degree of freedom
Only haplotypes with a frequency above 0.03 are shown

Table 3.3.3 illustrates the results of transmission disequilibrium testing (TDT) of the IL10 G and R haplotypes. The haplotype IL10.G12/R1 is shows a positive association with RA because the observed inheritance is greater than expected. This is in keeping with the TDT analysis of the IL10.R alleles. However, the result must be treated with caution because it is an uncommon haplotype and the numbers involved in the analysis are small.

Of greater importance, the haplotype IL10.G10/R3 shows a significant negative association with RA ($p < 0.005$) because it is inherited less frequently than expected. This negative association is in keeping with the results of the IL10.R TDT analysis but is stronger than that identified with IL10.R3 alone. This may indicate that the haplotype rather than the IL10.R3 allele is associated with RA. Although this haplotype is also uncommon the results can be viewed with a greater degree of confidence because the χ^2 result is greater than 5 [Spielman, 1996; Spielman, 1993].

Chapter 4

CHAPTER 4

Conclusions and future work

Rheumatoid arthritis (RA) is a chronic inflammatory disease predominantly involving synovial joints. It is common, affecting up to 1% of the population and is costly with a two-fold increased mortality, significant morbidity and high financial costs. Evidence from twin (Aho *et al.*, 1986; Silman *et al.*, 1993) and family (Deighton and Walker, 1991) studies suggest that both genetic and environmental factors contribute to susceptibility and disease heritability has been estimated to be approximately 60% (MacGregor *et al.*, 2000). The major histocompatibilty complex (MHC) on chromosome 6 has been consistently linked to and associated with RA susceptibility (Ollier and Thomson, 1992). However, family studies suggest that this accounts for only one-third of the genetic susceptibility (Deighton *et al.*, 1989; Wordsworth, 1991). It is likely that a number of other susceptibility loci exist, each contributing less than the MHC to the total genetic component.

Systematic linkage screening of the entire genome has the potential to detect all disease susceptibility loci if appropriately powered. However, it is necessary to study very large numbers of affected sibling pair (ASP) families if adequate power to detect linkage to non-MHC genes is to be generated (Brown and Wordsworth, 1998). Recruitment is difficult in a late onset disease such as RA and may require international collaboration. Undertaking whole genome analysis on such large numbers of individuals is both expensive and time consuming. Therefore the strategy has been to perform genome-wide screens on manageable numbers of families aiming to study any overlapping regions of nominal linkage ($p<0.05$) reported in two or more independent data sets in greater detail with larger numbers of independent ASPs (Holmans and Craddock, 1997).

Results from three whole genome scans (WGS) in Japanese (Shiozawa *et al.*, 1998), European (Cornelis *et al.*, 1998) and North American (Jawaheer *et al.*, 2001) RA families have already been reported and linkage to the MHC was confirmed ($p<2.5\times10^{-5}$) in the European and North American studies. A number of non-MHC regions with evidence of nominal linkage ($p<0.05$) were also identified. A number of these apparent linkages will be false positives and it is therefore important that all findings are independently confirmed in further data sets. Of the three reported WGS, nominal linkage to overlapping regions has been reported in two or more data sets for eight regions on chromosomes 3q, 8p, 12q, 14q, 16p, 16q, 18q and Xp (8, 9, 10) (Tables 2.2.1). These are therefore important candidate regions for containing true RA susceptibility loci, but, since the individual RA WGS have relatively low power to identify linkage to genes of small effect, some regions of true linkage are likely to have been missed.

Although linkage analysis is an integral part of the search for susceptibility genes it is both expensive and time consuming. The search for candidate genes, undertaken in parallel, can be a useful method of investigating biologically plausible candidate genes at considerably less cost in time and resources. Genes that control cytokine parameters are among the most obvious candidates to study in a chronic inflammatory disease such as rheumatoid arthritis.

Association studies are generally used to study candidate genes. The simplest form of association study is the case control study. This has the advantage of relatively easy collection of 'cases' as other family members are not required. However, non-random mating in a population leading to population stratification or assortative mating can produce spurious results. Family based association studies where the non-transmitted parental allele is used as the 'control' can overcome this problem. Most polymorphisms

studied are unlikely to be the real cause of disease susceptibility unless there is convincing evidence that the polymorphism is directly involved in influencing the production of a protein such as producing a stop codon and therefore a truncated protein. It is more likely that the polymorphism is a marker for other relevant mutations within the gene. Therefore, haplotyping, using a family based association study, is often undertaken to narrow down the region of interest.

This thesis presents results from both linkage and candidate gene studies designed to identify novel susceptibility loci, confirm linkage to previously identified regions and confirm or refute the involvement of IL10 as a gene involved in the susceptibility or severity of RA. This is a common and costly disease and learning more about factors predisposing to disease susceptibility and severity is without doubt of great importance.

SUMMARY OF LINKAGE STUDY RESULTS

Replication study

This study was designed to test for linkage to candidate susceptibility loci identified in the ECRAF study (Cornelis *et al.*, 1998), which was the first systematic linkage screen of the whole genome in RA. An independent cohort of 368 UK ASPs, fulfilling 1987 ACR criteria for RA (Arnett *et al.*, 1988), were used. Eighteen regions with nominal evidence of linkage ($p<0.05$) identified by single or multi-point analysis were studied using 59 suitable and informative microsatellites based within the regions of interest. Nominal evidence for linkage was identified for three markers by single point analysis: D12S95 ($p<0.05$), CYP19 ($p<0.05$) and D21S1252 ($p<0.02$). Multi-

point analysis, using MAPMAKER/SIBS, did not provide further evidence for linkage.

It was not surprising that only three of the 14 regions identified by the European genome screen were reproduced in the UK replication study. Linkage analyses undertaken in other complex diseases reveal the same lack of consistent results when studying independent family collections. Examples include studies of insulin dependent diabetes mellitus (IDDM) (Concannon *et al.*, 1998; Davies *et al.*, 1994; Luo *et al.*, 1996) and multiple sclerosis (Ebers *et al.*, 1996; Haines *et al.*, 1996). This inconsistency has been attributed to a number of factors including clinical and genetic heterogeneity, the use of different markers and analytical methods and poorly powered study designs unable to detect weak genetic effects. It is also accepted that a large proportion of apparent linkages identified by genome-wide linkage analysis will be false positives requiring confirmation of results in independent data sets (Lander and Kruglyak, 1995).

Two other regions identified by the European WGS, 18q22-23 and 3q13 contain the IDDM6 and IDDM9 loci and were studied in a further 194 European ASPs from 164 families (Cornelis *et al.*, 1998). The region 18q22-23 did not reach significance in the larger cohort. Suggestive linkage to the chromosome 3 region was found in the combined data set but was only evident in HLA concordant RA pairs (p=0.001). Linkage was not significant in the remaining RA pairs (p=0.08). Candidate genes within the 3q13 region include genes coding for CD80 and CD86 which interact with T cell molecules CTLA4 and CD28 in the co-stimulatory pathway (Cornelis *et al.*, 1998). No evidence for linkage in either region was identified by the UK replication study although the region on chromosome 3 at 3q13 may warrant further investigation as this area has been identified in two independent datasets (Cornelis *et al.*, 1998).

Whole genome linkage analysis

The UK genome-wide linkage analysis was initiated prior to the publication of the NARAC study and the aim was to generate additional regions where linkage may be found and to replicate any evidence of linkage at specific loci already identified. Once the North American WGS was reported nominal linkage to eight overlapping regions reported in two or more data sets (Cornelis *et al.*, 1998; Jawaheer *et al.*, 2001; Shiozawa *et al.*, 1998) was identified (table 2.2.1). These were important candidate regions for including true RA susceptibility loci. However, some regions of true linkage are likely to have been missed because most whole genome screens have relatively low power to identify linkage to genes of small effect (Brown and Wordsworth, 1998).

The UK WGS performed on a large well-characterised cohort of 182 UK Caucasian RA families and included 252 ASPs. The whole genome was spanned with a mean marker distance of ~ 10-cM using 368 highly informative microsatellites. Linkage to the HLA region on chromosome 6 was confirmed by multipoint analysis (max LOD = 4.7 at marker D6S276, p=0.000003, 1cM from HLA-DRB1). Suggestive evidence of linkage by multipoint analysis was identified on chromosome 6q nominal evidence of linkage by multipoint analysis was found for 3 additional linkage intervals; 1q, and two areas on 14q. Single point analysis also identified suggestive linkage for one marker on chromosome 6q and nominal linkage ($p<0.05$) was identified for 10 additional linkage intervals on chromosomes 3p, 4q, 7p, two sites on both 10q and 14q, 16p, 21q and Xq (table 2.3.6). Six of these non-HLA regions were also reported by the ECRAF (Cornelis *et al.*, 1998) and/or NARAC (Jawaheer *et al.*, 2001) studies (chromosomes 3q, 6q, 10q, 14p, 14q and 16p) and a number overlapped with regions linked to other autoimmune diseases (Tables 2.3.7 and 2.3.8). No supporting evidence was found for the IDDM 6 locus on chromosomes 18 or the IDDM 9 locus on chromosme 3,

sites which had shown evidence for linkage in the ECRAF study (Cornelis *et al.*, 1998).

A number of other studies have been undertaken using UK ASP families that overlap with those used in the UK WGS. John *et al* reported linkage to a single marker mapping to an intron within the Estrogen Synthase gene (CYP-19) (John *et al.*, 1999). Fife *et al* investigated Corticotrophin Releasing Hormone (CRH) as a candidate RA susceptibility locus and have reported both linkage and association at this site (Fife *et al.*, 2002; Fife *et al.*, 2000). The replication study discussed above identified three sites of nominal linkage by single point analysis: D12S95 ($p<0.05$), CYP19 ($p<0.05$) and D21S1252 ($p<0.02$). However, none of these regions were identified by the UK genome-wide linkage analysis. With the marker density used in this study (as with all WGS to date) linkage to these (and other postulated disease) loci could have been missed. Therefore, these regions cannot be excluded as potential disease genes.

Power calculations suggested that the WGS should have 80% power to detect loci contributing $\lambda = 1.6$ using the threshold for detection as a LOD $\geq$ 1.0. This may have been an overestimate because power calculations assume that all families are fully informative. Parental DNA was available in just over one third of the families but DNA from unaffected siblings was available in 94% of the remaining families. This was used to infer parental haplotypes when these were not directly obtainable and is one of the strengths of the study. However, HLA-DRB1 is thought to have a $\lambda = 1.6$ and the WGS (despite the fairly large numbers of ASPs studied) was calculated to have 80% power to detect loci contributing $\lambda = 1.6$ *i.e.* the equivalent to HLA. Non-HLA loci are thought to contribute to smaller λ values than HLA hence the power to detect such loci will be less. This may be an additional reason why a number of suspected regions were not identified.

In order to make best use of the information arising from this and other similar sized WGS, a number of strategies have been suggested. One proposal is to perform the initial screen in a relatively small data-set and then to test those loci demonstrating nominal evidence of linkage ($p<0.05$) in a larger data-set (Holmans and Craddock, 1997). Hence, the UK WGS could be seen as a replication data set in which to test positive linkages detected in the US and European whole genome studies. A complementary but alternative approach would be to undertake a meta-analysis analysis of the genotype data already obtained by published whole genome screens. By combining this data it should be possible to make more robust interpretations of the evidence so focusing the efforts of the replication studies.

SUMMARY OF CANDIDATE GENE STUDY RESULTS

IL10 case control studies

IL10 may be involved in susceptibility and/or severity to RA for a number of reasons. It is a cytokine with anti-inflammatory capabilities (Cassatella *et al.*, 1993; de Waal Malefyt *et al.*, 1991; Fiorentino *et al.*, 1989; Fiorentino *et al.*, 1991a; Katsikis *et al.*, 1994) but is a potent up-regulator of B cell production and differentiation (Rousset *et al.*, 1992). Transgenic mice with inactivated IL10 show normal development of lymphocytes and antibody responses but most animals are growth retarded, anaemic and suffer from chronic enterocolitis. Repeated administration of IL10 into these mutants can transiently cure their disease (Kuhn *et al.*, 1993). Variation in IL10 secretion is largely genetically determined (Westendorp *et al.*, 1997) and differences in secretion have been associated with various chronic inflammatory and infectious diseases (Derkx *et al.*, 1995; Kurtzhals *et al.*, 1998; Lehmann *et al.*, 1995; Westendorp *et al.*, 1997). The IL10 gene maps to the junction of 1q31-q32 (Eskdale *et al.*, 1997a). IL10 secretion has been associated with particular haplotypes defined by the microsatellite polymorphisms, IL10.G and

IL10.R (Eskdale and Gallagher, 1995; Eskdale *et al.*, 1996). Various genotypes of IL10.R, IL10.G and SNPs have been reported to show association with a variety of chronic inflammatory diseases (Crawley *et al.*, 1999; Mok *et al.*, 1998). The IL10.G microsatellite has been associated with SLE (Eskdale *et al.*, 1997b) and the IL10.R2 allele was over-represented and the IL10.R3 allele under-represented in rheumatoid arthritis (Eskdale *et al.*, 1998b).

The simplest form of association study is the case control study. It has the advantage of relatively easy collection of 'cases' as other family members are not required. In genetic studies, where only DNA is needed from the control population, normal blood donors can be used providing they are of the same ethnic background and sex match as the cases. To investigate whether an association could be identified between RA susceptibility and IL10.R microsatellite alleles two case control studies were undertaken; one in a Caucasian population consisting of individuals with severe RA and the other in a Black South African population not specifically selected for disease severity. No differences were observed in IL10.R allele frequencies between patients and controls in either racial group (table 3.2.2). The frequency of IL10.R2 was significantly higher ($p=1\times10^{-6}$) in the South African population overall (0.78) compared to the UK Caucasians (0.66) while the frequency of IL10.R3 was correspondingly reduced (0.16 versus 0.30, $p=1\times10^{-8}$). Subgroup analysis of the UK Caucasian RA cohort did not identify any association with IL10.R alleles.

These results were in contrast to previous reports of an increase in IL10.R2 and a reduction in IL10.R3 alleles (Eskdale *et al.*, 1998b). There are several possible explanations for this. Firstly, the previously reported association with the IL10.R2 allele may have been spurious. Secondly, it may have been relevant that the IL10.R2 allele frequency in UK Caucasian controls in the Oxford study were significantly higher (0.65 versus 0.59, p=0.03) than that

reported by Eskdale et al and the IL10.R3 frequency significantly lower (0.31 versus 0.38, p=0.02). This may suggest that IL10 contributes a weak genetic effect but the relatively high frequency of IL10.R2 in the general population made the effect more difficult to detect reliably. Since the frequency of IL10.R2 was even higher in South African Sotho and Zulu populations, the power to detect association with RA would be further reduced. Thirdly, genetic heterogeneity may be operating. The UK patients in this study were specifically selected for having more severe forms of RA and were only included if they had undergone an early large joint replacement or fulfilled the criteria for rheumatoid vasculitis or Felty's Syndrome. This was in contrast to the previous report where the main recruitment criteria was RA fulfilling the 1987 ACR criteria (Eskdale *et al.*, 1998b). It is therefore likely that the disease severity of the Caucasian patients included in the two studies was different and it is conceivable that the IL10.R2 allele is not as strongly associated with severe forms of RA as it is with milder variants.

IL10 family based association study

Non-random mating in a population leading to population stratification or assortative mating can produce spurious results in a case control study. Also, most polymorphisms studied are unlikely to be the real cause of disease susceptibility unless there is convincing evidence that the polymorphism is directly involved in influencing the production of a protein such as producing a stop codon and therefore a truncated protein. It is more likely that the polymorphism is a marker for other relevant mutations within the gene. Therefore, haplotyping is often undertaken to narrow down the region of interest (section 1.2.10.5) but this can only be performed using family based association methods e.g. TDT analysis.

A third study using TDT analysis was undertaken so that haplotyping of the IL10.R and IL10.G microsatellite markers in a cohort of UK Caucasians with

RA could be performed to determine whether specific haplotypes were associated with RA susceptibility. Families included 163 simplex Caucasian RA families recruited from UK; 55 from the Oxfordshire area and the remaining 108 from the ARC-UK National Repository of RA families.

TDT analysis using TRANSMIT (section 1.2.10.3) identified an association for IL10.R ($p<0.05$) but not for IL10.G. This was in keeping with the results of a previous case control study (Eskdale *et al.*, 1998b) but was in contrast to the results of the two case control studies discussed above. The IL10.R3 allele was under-represented in this rheumatoid population and excess transmission of the less common IL10.R1 allele was identified. The under-representation of IL10.R3 corresponded to an earlier study (Eskdale *et al.*, 1998b) but the over-representation of IL10.R1 had not been reported previously. Haplotype analysis revealed a stronger negative association with the haplotype IL10.R3/IL10.G10. This suggested that the locus IL10.R was not the primary association but that another polymorphism on the IL10.R3 / IL10.G10 haplotype may be involved.

The results of IL10 microsatellite association studies may vary because of differences in disease severity between the patient cohorts. A number of groups have reported associations between IL10 and disease severity rather than susceptibility (Mok *et al*, 1998; Huizinga *et al*, 2000). Although, the cohort in the study was not specifically chosen for disease severity, the individuals developed RA at a relatively young age and many of them had a family history of RA. Both early age of onset and family history are risk factors for more severe disease (Deighton *et al.*, 1994). Secondly, the high IL10.R2 allele frequency in the general population may make the detection of a weak genetic effect unreliable. The reported association with the IL10.R2 allele could be spurious and may be the effect of unidentified population stratification in one of the case control studies. A family based association study would have avoided this problem. Extended haplotyping in a family

based association study can narrow down the DNA sequence of interest by identifying strong associations between specific haplotypes and RA. Therefore, the stronger negative association with the haplotype IL10.R3/IL10.G10 suggests that the locus IL10.R is not the primary association and that another polymorphism on the IL10.R3/IL10.G10 haplotype may be involved. Extended haplotyping, using both microsatellites and SNPs in a further independent family dataset, would be the next step to confirm the association and narrow down the DNA sequence of interest.

IDENTIFYING THE PREDISPOSITION TO AND SEVERITY OF A DISEASE BY GENETIC STUDIES: THE PROBLEMS AND POSSIBILITIES FOR THE FUTURE

Genetic studies have the potential to help identify the causes of many difficult and complex diseases. The technology is in place and is advancing at great speed. However, the studies are hampered by a number of factors. The very nature of the work is complex and open to a considerable level of human error. This will continue to be problematic until further 'error proof' and affordable automated systems can be developed.

A number of statistical packages are being developed which attempt to extract the maximum power from a data set. However, these will not be used to the full until the diseases under study are more accurately defined (Roses, 2002b). Many of the complex diseases under investigation have no single specific test that can identify an individual sufferer. Most diagnoses rely on a series of clinical diagnostic criteria and therefore individuals included within the same study may have very different symptoms and signs. The long-term disease severity may also be very different. For instance, some individuals may not have RA if reviewed some years after the initial diagnosis as the original joint symptoms may have been due to a reactive arthritis which has

subsequently settled. Major advances in defining and understanding inherited and non-inherited factors contributing to RA may lead to the abandonment of the concept of RA as a single entity the acceptance of a heterogeneous model of the disease. Identifying distinct RA subsets and stratifying patients into categories that differ with respect to aetiology, disease course, clinical pattern and treatment response (Weyand and Goronzy, 1995) should improve the ability to identify the genes involved with disease predisposition. As understanding of the disease pathogenesis of RA improves new biological markers associated with RA may become apparent. For example recent research suggests that anti-cirtrulline antibodies could be used in conjunction with rheumatoid factor leading to a specificity for RA of 99.6% (Bizzaro *et al.*, 2001). Studies are ongoing to determine whether anti-cirtrulline antibodies could be used as a prognostic marker in conjunction with other factors.

Statistical programs assume 100% accuracy and availability of information. They do not account for unavailable parents leading to an inability to define identity by descent, PCR failures leading to a less than ideal success rate for a marker or series of markers. All these factors lead to a reduced level of power to identify the gene or region linked or associated to the disease. The non-paternity rate often quoted by genetic studies is in the region of 10 – 20%. Some of this must be in part due to a series of simple but common errors such as mislabelling a sample, placing DNA into the wrong row in a Costar plate, pipette contamination and other simple, well recognised laboratory errors.

Problems with ascertainment, diagnositic criteria, aetiological heterogeneity and failure to correct for the testing of multiple genetic models have all been blamed for the failure to replicate linkages in other family data sets (Risch, 1990a; Suarez and Hampe, 1994). To achieve replication in a complex trait caused by 6 loci with equal effects computer simulation has shown that

approximately five times the number of families needed for the first detection would be required (i.e. n-1, where n is the number of loci) (Davies *et al.*, 1994; Suarez *et al.*, 1994).

The human genome is vast and the information currently available is enormous and expanding. Microsatellite markers placed approximately 10-cM apart mean that some markers are within 10 million base pairs of each other but others may be nearer to 20 million base pairs away. Given these figures, it is not surprising that regions of possible linkage may be missed. Stringent correction for multiple testing is required when undertaking a WGS but because of this sites of potential linkage could be missed.

Technological advances will enable whole genome linkage studies using saturated bi-allelic markers to be undertaken. Already The SNP Consortium (TSC), a collaboration between ten pharmaceutical companies and the Wellcome Trust, have identified approximately 200,000 SNPs and placed them in the public domain (Group, 2001). The SNP assays have been tested and validated and can be used as the basis of a polymorphism map which could become a pharmaceutical industry standard (Roses, 2002b). An example of this approach is the mapping of APOE4 using a high density SNP map (Kruglyak, 1999; Martin *et al.*, 2000; Roses, 2000). APOE4 is the variant in the APOE gene associated with Alzheimer disease whereas APOE2 is protective (Roses, 2002a). Another example is in psoriasis with the construction of a SNP map over a 1.2 million base-pair (bp) linkage region. A small 80-100,00 bp region containing specific SNPs associated with psoriasis was localised (Hewett *et al.*, 2002). However, with hundreds of thousands of markers used the size of the correction will need to be very large.

Despite technological advances to generate adequate power to detect linkage to non-MHC genes it will be necessary to study very large numbers of affected sibling pair (ASP) families (Brown and Wordsworth, 1998). Alternative ways of genotyping these markers will be identified. DNA microarray technology is already being used as are robotics in some centres. Mixing of the DNA of 100 individuals into one tube will cut costs and speed results (Roses, 2002b). Although cost per SNP genotype in fully automated laboratories has fallen from $1.00 to $0.10 with the ultimate aim of reducing this cost to $0.01 per SNP genotype (Roses, 2002a). A genome-wide SNP screen will involve approximately 200,000 genotypes and cost reductions brought about by multiplexing and mixing techniques can reduce the cost per individual from $200,000 to $20,000 or possibly $2,000 per individual (Roses, 2002a). But even at $2,000 per person, with the large numbers of individuals to be studied this will be expensive and time consuming.

Recruitment remains difficult in a late onset disease such as RA and may require international collaboration. Time must be spent on careful study design and stringent diagnostic criteria for recruitment. One way forward would be to undertake case control studies using very large numbers of carefully identified individuals with very similar patterns of disease and genetic backgrounds. If extended haplotyping using SNPs had already been undertaken using family based methods, the haplotypes could be studied in the cases and control populations. This would have the advantage of avoiding problems of extensive multiple testing and the quantity of genotyping in relation to the number to affected individuals would be halved. However, these studies would only be able to investigate candidate genes and would not be able to generate additional regions where linkage may be found (i.e. unable to generate 'hypotheses'). Up to 80% of human genes remain unknown hence concentrating solely on candidate genes would reduce the possibilities for non-MHC detection. Even amongst identified genes,

candidates may not be considered until a linkage study identifies the region of the genome they reside in.

Candidate regions and genes can be inferred from investigations into other autoimmune diseases such as ankylosing spondylitis, insulin-dependent diabetes mellitis, inflammatory bowel disease and lupus erythematosus. This approach has been used for some time and studies have investigated the possibility of a set of 'autoimmune genes' common to all the autoimmune diseases such as cytokine genes (Myerscough *et al.*, 2000). Another area under study is that of animal models where synteny with animal chromosomes is investigated (Jacob *et al.*, 1992; Risch *et al.*, 1993). For instance candidate gene studies and genome-wide linkage studies have been undertaken in various forms of murine induced arthritis (Bergsteinsdottir *et al.*, 2000; Kawahito *et al.*, 1998; Otto *et al.*, 1999). As discussed in section 2.3.5 a WGS conducted using mice with proteoglygan induced arthritis identified a linkage interval (named pgia-3) homologous to a region identified by the UK genome-wide linkage study in RA and both the American and European genome screens (Cornelis *et al.*, 1998; Jawaheer *et al.*, 2001; Otto *et al.*, 1999). An area homologous to the TNF 2 receptor (TNFR II) on chromosome 17 was identified in mice and this has since been studied in RA families with interesting results (Barton *et al.*, 2001).

Another approach is to tackle the problem of disease heterogeneity by studying founder populations. Here the genetic background will be more homogeneous than for many data sets and the range of genes involved in a disease may be similar for all individuals. However, the results may not be fully representative of the general population. An example of this approach is in SLE where a region 2q37on chromosome 2 was identified within the Icelandic population (Lindqvist *et al.*, 2000).

Overall, as technology advances genetic studies will be undertaken with ever increasing complexity and greater numbers of genetic markers and DNA samples. Automation should reduce the chance of human error inherent in current studies so improving power. However, until it is easier to classify diseases, group individuals into selected cohorts based upon disease severity and identify their genetic backgrounds with greater care it will remain difficult to pick out the combinations of genes underlying the various complex diseases under study.

ADVANCES IN THERAPEUTICS AND POSSIBILITIES FOR THE FUTURE

Understanding the natural history of a disease at the molecular level, particularly in relation to the susceptibility to and severity of the disease, is important but unlikely to alter the incidence and prevalence of most multifactorial diseases for the foreseeable future. However, the advances in therapeutics forwarded by molecular biological techniques should (and already have) improve our ability to treat and control certain diseases. An example in RA is the recent introduction of biological therapies such as anti-TNF treatments which very successfully help to control disease activity in people with severe RA. These new therapies depend upon both the technological advances in molecular biology and a much broader understanding of the disease pathogenesis. Therapeutic avenues include recombinant DNA (rDNA) derived drugs, monoclonal antibodies (for diagnosis and treatment), vaccines, gene therapy including methods of gene delivery, modulating expression of genes, gene marking and the introduction of new genes.

Recombinant DNA technology has lead to the production of recombinant DNA (rDNA) derived drugs such as factors 8 and 9, erythropoietin, growth hormone and haematopoietic growth factors such as GM-CSF (granulocyte

Other individuals are sensitive to certain drugs as a result of differences in absorption, distribution, metabolism and elimination.

Pharmacogenetic differences between individuals can have profound effects on a drug's metabolism and efficacy. A genetic profile is difficult to change but if these differences were readily determined the drug dosage could be altered. An example of this approach in RA would be the measurement of the thiopurine methyltransferase (TMPT) enzyme prior to starting treatment with azathioprine (Corominas *et al.*, 2003; Marra *et al.*, 2002). Cytochrome P-450 enzymes form a family of ~30 liver-specific proteins involved in oxidative metabolism of a range of drugs and chemicals. The gene coding CYP2D6 is on chromosome 22 and is the best understood of the P-450 cytochrome enzymes. It is involved in the biotransformation of a wide range of drugs including anti-depressants, anti-arrythmics and proton pump inhibitors. Some individuals are slow and others fast metabolisers. Point mutations and deletions have been identified (slow metaboliser) and in a few individuals multiple copies of the gene have been present (fast metaboliser). Molecular analysis could, in the future, lead to drug metabolism status being assessed prior to starting any form of treatment thus avoiding toxic side effects. A example in the musculoskeletal clinic could be the individual taking warfarin whom is started on a proton pump inhibitor (e.g. omeprazole) to treat peptic ulceration (possibly as a result of other drug treatment). Omeprazole can induce the cytochrome P-450 system interfering with the action of warfarin.

An immediate impact of molecular biological techniques such as SNP technology will be in the area of pharmacogenetics. This is a discipline aiming to characterise a person with respect to disease susceptibility, severe adverse events (SAE) associated with taking a medicine or whether the medicine is effective for treatment or prevention of disease (Roses, 2002a). The generation of genetically associated targets for drug discovery in humans remains a goal but realistically even under the best circumstances of finding a

and macrophage colony stimulating factor). Previously individuals were treated with factors and hormones derived from pooled plasma, blood or pituitary extract obtained from cadavers. This lead to a number of iatrogenically acquired life threatening illnesses such as HIV infection, hepatitis C infection and Creutzfeldt-Jakob disease. rDNA derived drugs have meant that those newly diagnosed with conditions such as haemophilia are no longer at risk of developing another disease as a result of the treatment given to them for their primary condition. Other widely used rDNA derived drugs include human insulin and the recombinant hepatitis B virus vaccine. Work is underway to produce a recombinant HIV vaccine. In the future it may be possible to produce recombinant anti-inflammatory cytokines to treat RA. If a single infective cause is ever found to be the trigger for disease onset in a genetically susceptible individual a recombinant vaccine against that infective cause could be produced.

Antibodies are proteins made as a defence against foreign cells and organisms. Antibodies comprise two heavy and two light chains and the part that binds to the recognition sequence on the target (or antigen) is called the Fab portion. The remainder is called the Fc segment and defines the antibody's properties. Each antibody can recognise a single unique antigen and this ability is utilised in many diagnostic strategies. Antibodies are conventionally produced by immunising an animal (e.g. a rabbit) with an antigen but the antiserum produced contains a mixture of antibodies. This heterogeneity was overcome with the development of monoclonal antibodies. Monoclonal antibodies are used diagnostically in radioimmunoassays (RIAs), enzyme immunosorbent assays (EIAs), flow cytometry and in situ hybridisation for histological immunotyping of tissue sections. Monoclonal antibodies to tumour specific antigens can be linked to effector molecules such as drugs, radionucleotides or toxins thereby targeting treatment and minimising treatment side effects. IT is possible that such a method could be used in the future to target localised therapies to sites of chronic inflammation

such as an inflammed joint. Currently in RA and other inflammatory conditions (e.g. Crohns disease) recent advances have lead to the very successful use of anti-TNF treatments (e.g. infliximab (Maini *et al.*, 1999)) to control disease activity. These are mouse-derived monoclonal antibodies which have been 'humanised' as the treatments need to be given repetitively and a mouse-derived monoclonal antibody will have a short survival time in a human because it is foreign and will induce an immunological response against it's Fab and Fc segments. To humanise the antibody portions of the heavy and light chains are replaced with gene segments that are human in origin. The antigen recognition portion remains murine but the remainder including the Fc component is human. Some antigenicity is still directed at the Fab component hence these drugs are often used in conjunction with other immunosuppressents (e.g. infliximab and methotrexate are used together to treat RA).

Drug resistance remains a common problem in the treatments for cancer, infection but can also be a problem when treating inflammatory conditions such as RA where a small proportion of individuals initially respond to DMARDs but quickly become unresponsive to their effects. There are many ways that drug resistance can be influenced and include the route of administration, the concentration attained at the target site. Changes at the molecular level could include mutations that alter protein-binding affinity, increased gene expression via amplification or enhanced transcription and decreased uptake of the drug. One example of drug resistance in cancer is resistance to methotrexate where dihydrofolate reductase (DHFR) enzyme activity is increased as a result of gene amplification within the tumour cells. As the antimetabolite effect of methotrexate occurs through the inhibition of DHFR an increase of tumour cell DHFR expression negates this. A small proportion of patients with RA are also resistant to methotrexate even when the route of administration is changed from oral to subcutaneous therapy.

disease-specific gene target quickly, the drug development will take between 7 – 12 years hence (Roses, 2002b). Adverse event pharmacogenetics and efficacy pharmacogenetics however should produce results in the short to medium term. SNP profiles of individuals who develop side effects to drugs could be compared with those who are side effect free and a similar system could be applied to those who do not respond to a drug therapy. Ultimately a series of SNPs could be developed for each new drug to test a patient prior to starting treatment for both SAE and efficacy. The abbreviated informative SNP panels are called a Medicine Response Test or SNP Print™. These tests could be performed with current technologies, such as microarrays at reasonable cost (Roses, 2002a). This would have both safety and cost benefits by avoiding treatment of those individuals at risk of an adverse drug event or those for whom the medicine is ineffective. Approximately 30 – 40 per cent of patients with RA are unresponsive to anti-TNF therapies (Maini *et al.*, 1999; Weinblatt *et al.*, 2003; Weinblatt *et al.*, 1999). These are expensive treatments costing on average £10,000 per patient per year. Currently trial and error for a minimum of a three month period is the only way to determine efficacy. At approximately £1000 per person, costs mount for the 30 percent who fail treatment. Many other DMARDs used to treatment RA, although relatively inexpensive, are costly in terms of drug side effects (section 1.1.10). Any method of predicting these in individual patients prior to treatment would be of great benefit.

Gene therapy can be defined as the transfer of genetic material (DNA or RNA) into the cells of an organism. In humans it has involved somatic cell transfer, in animals it could also mean germline gene therapy. It can be used to alter the natural immunity of cells, to kill or interfere with the growth of cells, suppress oncogenes, insert genes and includes anti-sense technology. This therapy has the potential to treat genetic diseases such as immunodeficiencies (e.g. adenosine deaminase deficiency), cystic fibrosis or haemophilia and acquired disease such as cancer or AIDS. It also has the

potential to treat some multifactorial diseases such as those where inflammation is prominent. For example, a study in a SCID mouse transfer model has shown that colitis can be prevented by IL-10 transduced T lymphocytes (van Montfrans *et al.*, 2002). IL-10 plays a key role in maintaining the immune balance in the intestinal mucosa (in a similar manner to the anti-inflammatory role it plays in the joint). Primary CD4+ cells were engineered to express Il-10 and this expression protected against experimental colitis. It is possible a similar approach could be used in inflammatory arthritis. The IL-10 produced would counter-balance the high levels of pro-inflammatory cytokines and limit their production by leucocytes.

Molecular biological techniques will continue to play a major role in many areas of medical research and this potential will increase with each new advance. From the point of view of more common multi-factorial diseases affecting large numbers of individuals knowledge of disease pathogenesis is paramount if progress is to continue and the molecular techniques are to be used to their full advantage. In RA improved diagnostic criteria are required to determine exactly which patients are included in the different disease sets to be studied. If a wide range of patients with differing clinical signs are included in the same data-set it is possible that different diseases at a molecular level are being studied. This would reduce the power of any study to identify disease associations and ultimately reduce the likelihood of identifying successful drug therapies. Molecular biological techniques may lead to a diagnostic test for RA so patients could be categorised reliably (as in insulin-dependent diabetes mellitus). This should improve the outcome of both immunopathogenic and therapeutic studies. Until this is possible the way forward is to further define disease using clinical criteria and to take great care in the recruitment for such studies. Pharmacogenetics with the aim of avoiding adverse drug events and identifying those individuals who should respond to a drug therapy is likely to be another area of progress. This would be particularly valuable in chronic condition such as RA where many individuals either develop side effects or to do not respond drug treatments.

The advantages of avoiding adverse drug events are obvious as are the financial incentives, what is less obvious but of equal benefit in a progressive disease is identifying the 'correct' drug first time and so avoid the 'cost' of joint destruction that will continue until the disease is controlled. The BMJ in February 2002 may not have thought that RA was a genetic disease but many of the future advances in treatment are likely to come via this field of genetics.

References

REFERENCES

- Abecasis, G. R., Cardon, L. R., and Cookson, W. O., 2000, A general test of association for quantitative traits in nuclear families, *Am J Hum Genet* **66**(1):279-92.
- Aho, K., Koskenvuo, M., Tuominen, J., and Kaprio, J., 1986, Occurrence of rheumatoid arthritis in a nationwide series of twins., *J Rheumatol* **13**(5):899-902.
- Allison, D. B., 1997, Transmission-disequilibrium tests for quantitative traits, *Am J Hum Genet* **60**(3):676-90.
- Arend, W. P., 2002, The mode of action of cytokine inhibitors, *J Rheumatol Suppl* **65:**16-21.
- Arend, W. P., and Guthridge, C. J., 2000, Biological role of interleukin 1 receptor antagonist isoforms, *Ann Rheum Dis* **59 Suppl 1:**I60-4.
- Arnett, F. C., Edworthy, S. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., Healey, L. A., Kaplan, S. R., Liang, M. H., Luthra, H. S., and et al., 1988, The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis, *Arthritis Rheum* **31**(3):315-24.
- Bankhead, C., Silman, A., Barrett, B., Scott, D., and Symmons, D., 1996, Incidence of rheumatoid arthritis is not related to indicators of socioeconomic deprivation, *J Rheumatol* **23**(12):2039-42.
- Barrett, E. M., Scott, D. G., Wiles, N. J., Symmons, D. P., 2000, The impact of rheumatoid arthritis on employment status in the early years of disease: a UK community-based study, **39**:1403-1409
- Barton, A., John, S., Ollier, W. E., Silman, A., and Worthington, J., 2001, Association between rheumatoid arthritis and polymorphism of tumor necrosis factor receptor II, but not tumor necrosis factor receptor I, in Caucasians., *Arthritis Rheum* **44**(1):61-5.

- Beighton, P., Solomon, L., and Valkenburg, H. A., 1975, Rheumatoid arthritis in a rural South African Negro population, *Ann Rheum Dis* **34**(2):136-41.
- Bergsteinsdottir, K., Yang, H. T., Pettersson, U., and Holmdahl, R., 2000, Evidence for common autoimmune disease genes controlling onset, severity, and chronicity based on experimental models for multiple sclerosis and rheumatoid arthritis, *J Immunol* **164**(3):1564-8.
- Bizzaro, N., Mazzanti, G., Tonutti, E., Villalta, D., and Tozzoli, R., 2001, Diagnostic accuracy of the anti-citrulline antibody assay for rheumatoid arthritis., *Clin Chem* **47**(6):1089-93.
- Blower, A. L., Brooks, A., Fenn, G. C., Hill, A., Pearce, M. Y., Morant, S., and al., e., 1997, Emergency admissions for upper gastrointestinal disease and their relation to NSAID use., *Aliment Pharmacol Ther* **11**:283-91.
- Boehnke, M., and Cox, N. J., 1997, Accurate inference of relationships in sib-pair linkage studies, *Am J Hum Genet* **61**(2):423-9.
- Bower, A. C., 1998, Rheumatoid arthritis: imaging, in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby, London, pp. 5.5.1-5.5.8.
- Brown, M. A., Pile, K. D., Kennedy, L. G., Campbell, D., Andrew, L., March, R., Shatford, J. L., Weeks, D. E., Calin, A., and Wordsworth, B. P., 1998, A genome-wide screen for susceptibility loci in ankylosing spondylitis, *Arthritis Rheum* **41**(4):588-95.
- Brown, M. A., and Wordsworth, B. P., 1998, Genetic studies of common rheumatological diseases [editorial], *Br J Rheumatol* **37**(8):818-23.
- Bunce, M., Fanning, G. C., and Welsh, K. I., 1995, Comprehensive, serologically equivalent DNA typing for HLA-B by PCR using sequence-specific primers (PCR-SSP), *Tissue Antigens* **45**(2):81-90.
- Burdin, N., Peronne, C., Banchereau, J., and Rousset, F., 1993, Epstein-Barr virus transformation induces B lymphocytes to produce human interleukin 10, *J Exp Med* **177**(2):295-304.

- Cassatella, M. A., Meda, L., Bonora, S., Ceska, M., and Constantin, G., 1993, Interleukin 10 (IL-10) inhibits the release of proinflammatory cytokines from human polymorphonuclear leukocytes. Evidence for an autocrine role of tumor necrosis factor and IL-1 beta in mediating the production of IL-8 triggered by lipopolysaccharide, *J Exp Med* **178**(6):2207-11.
- Chomarat, P., Vannier, E., Dechanet, J., Rissoan, M. C., Banchereau, J., Dinarello, C. A., and Miossec, P., 1995, Balance of IL-1 receptor antagonist/IL-1 beta in rheumatoid synovium and its regulation by IL-4 and IL-10, *J Immunol* **154**(3):1432-9.
- Chu, C. Q., Field, M., Allard, S., Abney, E., Feldmann, M., and Maini, R. N., 1992, Detection of cytokines at the cartilage/pannus junction in patients with rheumatoid arthritis: implications for the role of cytokines in cartilage destruction and repair, *Br J Rheumatol* **31**(10):653-61.
- Concannon, P., Gogolin-Ewens, K. J., Hinds, D. A., Wapelhorst, B., Morrison, V. A., Stirling, B., Mitra, M., Farmer, J., Williams, S. R., Cox, N. J., Bell, G. I., Risch, N., and Spielman, R. S., 1998, A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus [see comments], *Nat Genet* **19**(3):292-6.
- Consortium, I. H. G. S., 2001, Initial sequencing and analysis of the human genome., *Nature* **409**:860-921.
- Cooper N. J, Mugford, M., Scott, D. G. I., Barrett, E. M., Symmons, D. P., 2000, Secondary health service care and second line drug costs of early inflammatory polyarthritis in Norfolk, UK, J Rheumatol **27**:2115-22
- Cooper N. J, Mugford, M., Symmons, D. P., Barrett, E. M., Scott, D. G. I., 2002, Total costs and predictors of costs in individuals with early inflammatory polyarthritis: a community-based prospective study, Rheumatology **41**:767-774
- Cordell, H. J., Kawaguchi, Y., Todd, J. A., and Farrall, M., 1995, An extension of the Maximum Lod Score method to X-linked loci, *Ann Hum Genet* **59**(Pt 4):435-49.

- Cornelis, F., Faure, S., Martinez, M., Prud'homme, J. F., Fritz, P., Dib, C., Alves, H., Barrera, P., de Vries, N., Balsa, A., Pascual-Salcedo, D., Maenaut, K., Westhovens, R., Migliorini, P., Tran, T. H., Delaye, A., Prince, N., Lefevre, C., Thomas, G., Poirier, M., Soubigou, S., Alibert, O., Lasbleiz, S., Fouix, S., Weissenbach, J., and et al., 1998, New susceptibility locus for rheumatoid arthritis suggested by a genome- wide linkage study, *Proc Natl Acad Sci U S A* **95**(18):10746-50.
- Corominas, H., Domenech, M., Laiz, A., Gich, I., Geli, C., Diaz, C., de Cuevillas, F., Moreno, M., Vazquez, G., and Baiget, M., 2003, Is thiopurine methyltransferase genetic polymorphism a major factor for withdrawal of azathioprine in rheumatoid arthritis patients?, *Rheumatology* **42**(1):40-5.
- Cox, A., Camp, N. J., Cannings, C., di Giovine, F. S., Dale, M., Worthington, J., John, S., Ollier, W. E., Silman, A. J., and Duff, G. W., 1999a, Combined sib-TDT and TDT provide evidence for linkage of the interleukin-1 gene cluster to erosive rheumatoid arthritis, *Hum Mol Genet* **8**(9):1707-13.
- Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I., and Kong, A., 1999b, Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans, *Nat Genet* **21**(2):213-5.
- Crawley, E., Kay, R., Sillibourne, J., Patel, P., Hutchinson, I., and Woo, P., 1999, Polymorphic haplotypes of the interleukin-10 5' flanking region determine variable interleukin-10 transcription and are associated with particular phenotypes of juvenile rheumatoid arthritis, *Arthritis Rheum* **42**(6):1101-8.
- Curtis, D., 1997, Use of siblings as controls in case-control association studies, *Ann Hum Genet* **61**(Pt 4):319-33.
- D'Alfonso, S., Rampi, M., Rolando, V., Giordano, M., and Momigliano-Richiardi, P., 2000, New polymorphisms in the IL-10 promoter region., *Genes and immunity* **1**:213-233.

- Davies, J. L., Kawaguchi, Y., Bennett, S. T., Copeman, J. B., Cordell, H. J., Pritchard, L. E., Reed, P. W., Gough, S. C., Jenkins, S. C., Palmer, S. M., and et al., 1994, A genome-wide search for human type 1 diabetes susceptibility genes, *Nature* **371**(6493):130-6.
- de Groote, and Zangerel, 1992, *Cytokine* **4:**239-48.
- de Waal Malefyt, R., Abrams, J., Bennett, B., Figdor, C. G., and de Vries, J. E., 1991, Interleukin 10(IL-10) inhibits cytokine synthesis by human monocytes: an autoregulatory role of IL-10 produced by monocytes, *J Exp Med* **174**(5):1209-20.
- de Waal Malefyt, R., Yssel, H., Roncarolo, M. G., Spits, H., and de Vries, J. E., 1992, Interleukin-10, *Curr Opin Immunol* **4**(3):314-20.
- Deighton, C., Heslop, P., McDonagh, J., Walker, D., and Thomson, G., 1994, Does genetic anticipation occur in familial rheumatoid arthritis?, *Ann Rheum Dis* **53**(12):833-5.
- Deighton, C. M., and Walker, D. J., 1991, The familial nature of rheumatoid arthritis, *Ann Rheum Dis* **50**(1):62-5.
- Deighton, C. M., Walker, D. J., Griffiths, I. D., and Roberts, D. F., 1989, The contribution of HLA to rheumatoid arthritis, *Clin Genet* **36**(3):178-82.
- den Broeder, A. A., Joosten, L. A., Saxne, T., Heinegard, D., Fenner, H., Miltenburg, A. M., Frasa, W. L., van Tits, L. J., Buurman, W. A., van Riel, P. L., van de Putte, L. B., and Barrera, P., 2002, Long term anti-tumour necrosis factor alpha monotherapy in rheumatoid arthritis: effect on radiological course and prognostic value of markers of cartilage turnover and endothelial activation., *Ann Rheum Dis* **61**(4):311-8.
- Derkx, B., Marchant, A., Goldman, M., Bijlmer, R., and van Deventer, S., 1995, High levels of interleukin-10 during the initial phase of fulminant meningococcal septic shock, *J Infect Dis* **171**(1):229-32.
- Doherty, D. G., and Nepom, G. T., 1998, The human major histocompatibility complex and disease susceptibility, in: *Principles and practice of medical genetics* (D. L. Rimon, J. M. Connor, and R. E. Pyeritz, eds.), pp. 479-504.

- Drossaers-Bakker, K. W., Zwinderman, A. H., Vlieland, T. P., Van Zeben, D., Vos, K., Breedveld, F. C., and Hazes, J. M., 2002, Long-term outcome in rheumatoid arthritis: a simple algorithm of baseline parameters can predict radiographic damage, disability, and disease course at 12-year followup, *Arthritis Rheum* **47**(4):383-90.
- Dugowson, C. E., Koepsell, T. D., Voigt, L. F., Bley, L., Nelson, J. L., and Daling, J. R., 1991, Rheumatoid arthritis in women. Incidence rates in group health cooperative, Seattle, Washington, 1987-1989, *Arthritis Rheum* **34**(12):1502-7.
- Eberhardt, K., Fex, E., Johnson, U., and Wollheim, F. A., 1996, Associations of HLA-DRB and -DQB genes with two and five year outcome in rheumatoid arthritis, *Ann Rheum Dis* **55**(1):34-9.
- Ebers, G. C., Kukay, K., Bulman, D. E., Sadovnick, A. D., Rice, G., Anderson, C., Armstrong, H., Cousin, K., Bell, R. B., Hader, W., Paty, D. W., Hashimoto, S., Oger, J., Duquette, P., Warren, S., Gray, T., O'Connor, P., Nath, A., Auty, A., Metz, L., Francis, G., Paulseth, J. E., Murray, T. J., Pryse-Phillips, W., Risch, N., and et al., 1996, A full genome search in multiple sclerosis, *Nat Genet* **13**(4):472-6.
- Edwards, J. C. W., 1998, Rheumatoid arthritis: the synovium., in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby, London, pp. 5.6.1-5.6.8.
- Ehm, M., and Wagner, M., 1998, A test statistic to detect errors in sib-pair relationships, *Am J Hum Genet* **62**(1):181-8.
- Emery, P., and Salmon, M., 1995, Early rheumatoid arthritis: time to aim for remission?, *Ann Rheum Dis* **54**(12):944-7.
- Eskdale, J., and Gallagher, G., 1995, A polymorphic dinucleotide repeat in the human IL-10 promoter, *Immunogenetics* **42**(5):444-5.
- Eskdale, J., Gallagher, G., Verweij, C. L., Keijsers, V., Westendorp, R. G., and Huizinga, T. W., 1998a, Interleukin 10 secretion in relation to human IL-10 locus haplotypes, *Proc Natl Acad Sci U S A* **95**(16):9465-70.

- Eskdale, J., Keijsers, V., Huizinga, T., and Gallagher, G., 1999, Microsatellite alleles and single nucleotide polymorphisms (SNP) combine to form four major haplotype families at the human interleukin-10 (IL-10) locus, *Genes and Immunity* **1**:151-155.
- Eskdale, J., Kube, D., and Gallagher, G., 1996, A second polymorphic dinucleotide repeat in the 5' flanking region of the human IL10 gene, *Immunogenetics* **45**(1):82-3.
- Eskdale, J., Kube, D., Tesch, H., and Gallagher, G., 1997a, Mapping of the human IL10 gene and further characterization of the 5' flanking sequence, *Immunogenetics* **46**(2):120-8.
- Eskdale, J., McNicholl, J., Wordsworth, P., Jonas, B., Huizinga, T., Field, M., and Gallagher, G., 1998b, Interleukin-10 microsatellite polymorphisms and IL-10 locus alleles in rheumatoid arthritis susceptibility [letter], *Lancet* **352**(9136):1282-3.
- Eskdale, J., Wordsworth, P., Bowman, S., Field, M., and Gallagher, G., 1997b, Association between polymorphisms at the human IL-10 locus and systemic lupus erythematosus [published erratum appears in Tissue Antigens 1997 Dec;50(6):699], *Tissue Antigens* **49**(6):635-9.
- Fei, B. Y., Deng, C. S., Xia, B., Zhu, Y. Q., Crusius, J. B., and Pena, A. S., 2002, Polymorphisms at the TNF locus in Chinese Han population, *Hum Immunol* **63**(1):71-5.
- Felts, W., and Yelin, E., 1989, The economic impact of the rheumatic diseases in the United States, *J Rheumatol* **16**(7):867-84.
- Fife, M., Steer, S., Fisher, S., Newton, J., McKay, K., Worthington, J., Shah, C., Polley, A., Rosenthal, A., Ollier, W., Lewis, C., Wordsworth, P., and Lanchbury, J., 2002, Association of familial and sporadic rheumatoid arthritis with a single corticotropin-releasing hormone genomic region (8q12.3) haplotype, *Arthritis Rheum* **46**(1):75-82.
- Fife, M. S., Fisher, S. A., John, S., Worthington, J., Shah, C. J., Ollier, W. E., Panayi, G. S., Lewis, C. M., and Lanchbury, J. S., 2000, Multipoint linkage analysis of a candidate gene locus in rheumatoid arthritis

demonstrates significant evidence of linkage and association with the corticotropin-releasing hormone genomic region, *Arthritis Rheum* **43**(8):1673-8.

- Fiorentino, D. F., Bond, M. W., and Mosmann, T. R., 1989, Two types of mouse T helper cell. IV. Th2 clones secrete a factor that inhibits cytokine production by Th1 clones, *J Exp Med* **170**(6):2081-95.
- Fiorentino, D. F., Zlotnik, A., Mosmann, T. R., Howard, M., and O'Garra, A., 1991a, IL-10 inhibits cytokine production by activated macrophages, *J Immunol* **147**(11):3815-22.
- Fiorentino, D. F., Zlotnik, A., Vieira, P., Mosmann, T. R., Howard, M., Moore, K. W., and O'Garra, A., 1991b, IL-10 acts on the antigen-presenting cell to inhibit cytokine production by Th1 cells, *J Immunol* **146**(10):3444-51.
- Firestein, G. S., 1998, Rheumatoid synovium and pannus., in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby, London, pp. 5.13.1-5.13.24.
- Firestein, G. S., Paine, M. M., and Boyle, D. L., 1994, Mechanisms of methotrexate action in rheumatoid arthritis. Selective decrease in synovial collagenase gene expression., *Arthritis Rheum* **37**(2):193-200.
- Firestein, G. S., Paine, M. M., and Littman, B. H., 1991, Gene expression (collagenase, tissue inhibitor of metalloproteinases, complement, and HLA-DR) in rheumatoid arthritis and osteoarthritis synovium. Quantitative analysis and effect of intraarticular corticosteroids., *Arthritis Rheum* **34**(9):1094-105.
- Fulker, D. W., and Cardon, L. R., 1994, A sib-pair approach to interval mapping of quantitative trait loci., *Am J Hum Genet* **54**(6):1092-103.
- Gabriel, S. E., Crowson, C. S., Kremers, H. M., Doran, M. F., Turesson, C., O'Fallon, W. M., and Matteson, E. L., 2003, Survival in rheumatoid arthritis: a population-based analysis of trends over 40 years., *Arthritis Rheum* **48**(1):54-8.

- Gallagher, G., Eskdale, J., Oh, H. H., Richards, S. D., Campbell, D. A., and Field, M., 1997, Polymorphisms in the TNF gene cluster and MHC serotypes in the West of Scotland, *Immunogenetics* **45**(3):188-94.
- Garrod, A. B., 1859, The nature and treatment of gout and rheumatoid gout., Walton and Maberley, London.
- Goemaere, S., Ackerman, C., Goethals, K., De Keyser, F., Van der Straeten, C., Verbruggen, G., Mielants, H., and Veys, E. M., 1990, Onset of symptoms of rheumatoid arthritis in relation to age, sex and menopausal transition, *J Rheumatol* **17**(12):1620-2.
- Goodenow, M., Huet, T., Saurin, W., Kwok, S., Sninsky, J., and Wain-Hobson, S., 1989, HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions, *J Acquir Immune Defic Syndr* **2**(4):344-52.
- Gordon, D. A., and Hastings, D. E., 1998, Rheumatoid arthritis: clinical features of early, progressive and late disease, in: *Rheumatology* (J. H. Klippel, and P. A. Dieppe, eds.), Mosby, London, pp. 5.3.1-5.3.14.
- Goring, H. H., and Ott, J., 1997, Relationship estimation in affected sib pair analysis of late-onset diseases, *Eur J Hum Genet* **5**(2):69-77.
- Gough, A., Faint, J., Salmon, M., and al, e., 1994, Genetic typing of patients with inflammatory arthritis at presentation can be used to predict outcome., *Arthritis Rheum* **37**:1166-70.
- Greenwood, B., 1968, Autoimmune disease and parasitic infections in Nigerians., *Lancet* **2**:380-382.
- Gregersen, P. K., Silver, J., and Winchester, R. J., 1987, The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis, *Arthritis Rheum* **30**(11):1205-13.
- Grennan, D. M., and Sanders, P. A., 1988, Rheumatoid arthritis, *Baillieres Clin Rheumatol* **2**(3):585-601.
- Group, T. I. S. M. W., 2001, A map of the human genome sequence variation containing 1.42 million single nucleotide polymorphisms., *Nature* **409**:928-933.

- Haines, J. L., Ter-Minassian, M., Bazyk, A., Gusella, J. F., Kim, D. J., Terwedow, H., Pericak-Vance, M. A., Rimmler, J. B., Haynes, C. S., Roses, A. D., Lee, A., Shaner, B., Menold, M., Seboun, E., Fitoussi, R. P., Gartioux, C., Reyes, C., Ribierre, F., Gyapay, G., Weissenbach, J., Hauser, S. L., Goodkin, D. E., Lincoln, R., Usuku, K., Oksenberg, J. R., and et al., 1996, A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex. The Multiple Sclerosis Genetics Group, *Nat Genet* **13**(4):469-71.
- Hajeer, A. H., MacGregor, A. J., Rigby, A. S., Ollier, W. E., Carthy, D., and Silman, A. J., 1994, Influence of previous exposure to human parvovirus B19 infection in explaining susceptibility to rheumatoid arthritis: an analysis of disease discordant twin pairs, *Ann Rheum Dis* **53**(2):137-9.
- Halberg, P., 1998, Rheumatoid arthritis: history., in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby International, London, pp. 5.1.1.
- Hardingham, T. E., and Fosang, A. J., 1992, Proteoglycans: many forms and many functions., *FASEB J* **6**(3):861-70.
- Harrison, B., Symmons, D., Barrett, E., Silman, A., 1998, THe performance of the 1987 ARA classification criteria for rheumatoid arthritis ina population based cohort of patients with early inflammatory polyarthritis, J Rheumatol, **42**:1339-1346
- Harrison, B., Thomson, W., Symmons, D., Ollier, B., Wiles, N., Payton, T., Barrett, E., and Silman, A., 1999, The influence of HLA-DRB1 alleles and rheumatoid factor on disease outcome in an inception cohort of patients with early inflammatory arthritis, *Arthritis Rheum* **42**(10):2174-83.
- Hashimoto, L., Habita, C., Beressi, J. P., Delepine, M., Besse, C., Cambon-Thomsen, A., Deschamps, I., Rotter, J. I., Djoulah, S., James, M. R., and et al., 1994, Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q, *Nature* **371**(6493):161-4.

- Hauser, E. R., Boehnke, M., Guo, S. W., and Risch, N., 1996, Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations., *Genet Epidemiol* **13**(2):117-37.
- Heliovaara, M., Aho, K., Reunanen, A., Knekt, P., and Aromaa, A., 1995, Parity and risk of rheumatoid arthritis in Finnish women, *Br J Rheumatol* **34**(7):625-8.
- Hewett, D., Samuelsson, L., Polding, J., Enlund, F., Smart, D., Cantone, K., See, C. G., Chadha, S., Inerot, A., Enerback, C., Montgomery, D., Christodolou, C., Robinson, P., Matthews, P., Plumpton, M., Wahlstrom, J., Swanbeck, G., Martinsson, T., Roses, A., Riley, J., and Purvis, I., 2002, Identification of a psoriasis susceptibility candidate gene by linkage disequilibrium mapping with a localized single nucleotide polymorphism map., *Genomics* **79**(3):305-14.
- Holmans, P., and Craddock, N., 1997, Efficient strategies for genome scanning using maximum-likelihood affected-sib-pair analysis [see comments], *Am J Hum Genet* **60**(3):657-66.
- Horwitz, M., 2000, Basic concepts in medical genetics: a student's survival guide., McGraw-Hill, pp. p106.
- Huizinga, T. W., Keijsers, V., Yanni, G., Hall, M., Ramage, W., Lanchbury, J., Pitzalis, C., Drossaers-Bakker, W. K., Westendorp, R. G., Breedveld, F. C., Panayi, G., and Verweij, C. L., 2000, Are differences in interleukin 10 production associated with joint damage? [In Process Citation], *Rheumatology (Oxford)* **39**(11):1180-8.
- Ishizuka, T., Okayama, Y., Kobayashi, H., and Mori, M., 1999, Interleukin-10 is localized to and released by human lung mast cells, *Clin Exp Allergy* **29**(10):1424-32.
- Jacob, H. J., Pettersson, A., Wilson, D., Mao, Y., Lernmark, A., and Lander, E. S., 1992, Genetic dissection of autoimmune type I diabetes in the BB rat, *Nat Genet* **2**(1):56-60.
- Jacobsson, L. T., Hanson, R. L., Knowler, W. C., Pillemer, S., Pettitt, D. J., McCance, D. R., and Bennett, P. H., 1994, Decreasing incidence and

prevalence of rheumatoid arthritis in Pima Indians over a twenty-five-year period., *Arthritis Rheum* **37**(8):1158-65.

- Jacquard, A., 1978, Genetics of HUman Populations, Freeman, Cooper & Co., San Francisco.
- Jawaheer, D., Seldin, M. F., Amos, C. I., Chen, W. V., Shigeta, R., Monteiro, J., Kern, M., Criswell, L. A., Albani, S., Nelson, J. L., Clegg, D. O., Pope, R., Schroeder, H. W., Jr., Bridges, S. L., Jr., Pisetsky, D. S., Ward, R., Kastner, D. L., Wilder, R. L., Pincus, T., Callahan, L. F., Flemming, D., Wener, M. H., and Gregersen, P. K., 2001, A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases, *Am J Hum Genet* **68**(4):927-36.
- Jeurissen, M. E., Boerbooms, A. M., van de Putte, L. B., Doesburg, W. H., and Lemmens, A. M., 1991, Influence of methotrexate and azathioprine on radiologic progression in rheumatoid arthritis. A randomized, double-blind study, *Ann Intern Med* **114**(12):999-1004.
- John, S., Myerscough, A., Eyre, S., Roby, P., Hajeer, A., Silman, A. J., Ollier, W. E., and Worthington, J., 1999, Linkage of a marker in intron D of the estrogen synthase locus to rheumatoid arthritis, *Arthritis Rheum* **42**(8):1617-20.
- Kaplan, N. L., Martin, E. R., and Weir, B. S., 1997, Power studies for the transmission/disequilibrium tests with multiple alleles, *Am J Hum Genet* **60**(3):691-702.
- Katsikis, P. D., Chu, C. Q., Brennan, F. M., Maini, R. N., and Feldmann, M., 1994, Immunoregulatory role of interleukin 10 in rheumatoid arthritis, *J Exp Med* **179**(5):1517-27.
- Kawahito, Y., Cannon, G. W., Gulko, P. S., Remmers, E. F., Longman, R. E., Reese, V. R., Wang, J., Griffiths, M. M., and Wilder, R. L., 1998, Localization of quantitative trait loci regulating adjuvant-induced arthritis in rats: evidence for genetic factors common to multiple autoimmune diseases., *J Immunol* **161**(8):4411-9.

- Keystone, E., Wherry, J., and Grint, P., 1998, IL-10 as a therapeutic strategy in the treatment of rheumatoid arthritis, *Rheum Dis Clin North Am* **24**(3):629-39.
- Kim, J. M., Brannan, C. I., Copeland, N. G., Jenkins, N. A., Khan, T. A., and Moore, K. W., 1992, Structure of the mouse IL-10 gene and chromosomal localization of the mouse and human genes., *J Immunol* **148**(11):3618-23.
- Kirwan, J. R., 1995, The effect of glucocorticoids on joint destruction in rheumatoid arthritis. The Arthritis and Rheumatism Council Low-Dose Glucocorticoid Study Group, *N Engl J Med* **333**(3):142-6.
- Kong, A., and Cox, N. J., 1997, Allele-sharing models: LOD scores and accurate linkage tests, *Am J Hum Genet* **61**(5):1179-88.
- Kouri, T., Petersen, J., Rhodes, G., Aho, K., Palosuo, T., Heliovaara, M., Isomaki, H., von Essen, R., and Vaughan, J. H., 1990, Antibodies to synthetic peptides from Epstein-Barr nuclear antigen-1 in sera of patients with early rheumatoid arthritis and in preillness sera, *J Rheumatol* **17**(11):1442-9.
- Kruglyak, L., 1999, Prospects for whole-genome linkage disequilibrium mapping of common disease genes., *Nature Genetics* **22**(2):139-44.
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S., 1996, Parametric and nonparametric linkage analysis: a unified multipoint approach, *Am J Hum Genet* **58**(6):1347-63.
- Kruglyak, L., and Lander, E. S., 1995, Complete multipoint sib-pair analysis of qualitative and quantitative traits [see comments], *Am J Hum Genet* **57**(2):439-54.
- Kuhn, R., Lohler, J., Rennick, D., Rajewsky, K., and Muller, W., 1993, Interleukin-10-deficient mice develop chronic enterocolitis [see comments], *Cell* **75**(2):263-74.
- Kurtzhals, J. A., Adabayeri, V., Goka, B. Q., Akanmori, B. D., Oliver-Commey, J. O., Nkrumah, F. K., Behr, C., and Hviid, L., 1998, Low plasma concentrations of interleukin 10 in severe malarial anaemia compared with

cerebral and uncomplicated malaria [published errata appear in Lancet 1998 Jul 18;352(9123):242 and 1999 Mar 6;353(9155):848], *Lancet* **351**(9118):1768-72.

- Lander, E., and Kruglyak, L., 1995, Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results [see comments], *Nat Genet* **11**(3):241-7.
- Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., and Newburg, L., 1987, MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations, *Genomics* **1**(2):174-81.
- Lander, E. S., and Schork, N. J., 1994, Genetic dissection of complex traits, *Science* **265**(5181):2037-48.
- Laval, S. H., Timms, A., Edwards, S., Bradbury, L., Brophy, S., Milicic, A., Rubin, L., Siminovitch, K. A., Weeks, D. E., Calin, A., Wordsworth, B. P., and Brown, M. A., 2001, Whole-genome screening in ankylosing spondylitis: evidence of non-MHC genetic-susceptibility loci, *Am J Hum Genet* **68**(4):918-26.
- Lawrence, J. S., 1970, Rheumatoid arthritis - nature or nurture?, *Ann Rheum Dis* **29**:357.
- Leal, S. M., and Ott, J., 2000, Effects of stratification in the analysis of affected-sib-pair data: benefits and costs, *Am J Hum Genet* **66**(2):567-75.
- Lehmann, A. K., Halstensen, A., Sornes, S., Rokke, O., and Waage, A., 1995, High levels of interleukin 10 in serum are associated with fatality in meningococcal disease, *Infect Immun* **63**(6):2109-12.
- Lindqvist, A. K., Steinsson, K., Johanneson, B., Kristjansdottir, H., Arnasson, A., Grondal, G., Jonasson, I., Magnusson, V., Sturfelt, G., Truedsson, L., Svenungsson, E., Lundberg, I., Terwilliger, J. D., Gyllensten, U. B., and Alarcon-Riquelme, M. E., 2000, A susceptibility locus for human systemic lupus erythematosus (hSLE1) on chromosome 2q, *J Autoimmun* **14**(2):169-78.

- Linos, A., Worthington, J. W., O'Fallon, W. M., and Kurland, L. T., 1980, The epidemiology of rheumatoid arthritis in Rochester, Minnesota: a study of incidence, prevalence, and mortality., *Am J Epidemiol* **111**(1):87-98.
- Lipsky, P. E., van der Heijde, D. M., St Clair, E. W., Furst, D. E., Breedveld, F. C., Kalden, J. R., Smolen, J. S., Weisman, M., Emery, P., Feldmann, M., Harriman, G. R., and Maini, R. N., 2000, Infliximab and Methotrexate in the Treatment of Rheumatoid Arthritis, *N Engl J Med* **343**(22):1594-1602.
- Luo, D. F., Buzzetti, R., Rotter, J. I., Maclaren, N. K., Raffel, L. J., Nistico, L., Giovannini, C., Pozzilli, P., Thomson, G., and She, J. X., 1996, Confirmation of three susceptibility genes to insulin-dependent diabetes mellitus: IDDM4, IDDM5 and IDDM8, *Hum Mol Genet* **5**(5):693-8.
- Lynn, A. H., Kwoh, C. K., Venglish, C. M., Aston, C. e., and Chakravarti, A., 1995, Genetic epidemiology of rheumatoid arthritis., *Am. J. Hum. Genet.* **57**:150-59.
- MacGregor, A., Ollier, W., Thomson, W., Jawaheer, D., and Silman, A., 1995, HLA-DRB1*0401/0404 genotype and rheumatoid arthritis: increased association in men, young age at onset, and disease severity, *J Rheumatol* **22**(6):1032-6.
- MacGregor, A. J., 1995, Classification criteria for rheumatoid arthritis, *Baillieres Clin Rheumatol* **9**(2):287-304.
- MacGregor, A. J., Bamber, S., and Silman, A. J., 1994a, A comparison of the performance of different methods of disease classification for rheumatoid arthritis. Results of an analysis from a nationwide twin study, *J Rheumatol* **21**(8):1420-6.
- MacGregor, A. J., Riste, L. K., Hazes, J. M., and Silman, A. J., 1994b, Low prevalence of rheumatoid arthritis in black-Caribbeans compared with whites in inner city Manchester [see comments], *Ann Rheum Dis* **53**(5):293-7.

- MacGregor, A. J., and Silman, A. J., 1998, Classification and epidemiology, in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby, London, pp. 5.2.1-5.2.6.
- MacGregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K., and Silman, A. J., 2000, Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins., *Arthritis Rheum* **43**(1):30-7.
- MacKay, K., Eyre, S., Barrett, J., and al, e., 2000, Linkage analysis of potential rheumatoid arthritis non-HLA susceptibility loci, *Arthritis Rheum* **43**(9 (Suppl)):18.
- MacKay, K., Eyre, S., Myerscough, A., Milicic, A., Barton, A., Laval, S., Barrett, J., Lee, D., White, S., John, S., Brown, M. A., Bell, J., Silman, A., Ollier, W., Wordsworth, P., and Worthington, J., 2002, Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom., *Arthritis Rheum* **46**(3):632-9.
- Maini, R., St Clair, E. W., Breedveld, F., Furst, D., Kalden, J., Weisman, M., Smolen, J., Emery, P., Harriman, G., Feldmann, M., and Lipsky, P., 1999, Infliximab (chimeric anti-tumour necrosis factor alpha monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial. ATTRACT Study Group., *Lancet* **354**:1932-9.
- Marra, C. A., Esdaile, J. M., and Anis, A. H., 2002, Practical Pharmacogenetics: The Cost Effectiveness of Screening for Thiopurine s-Methyltransferase Polymorphisms in Patients with Rheumatological Conditions Treated with Azathioprine., *J Rheumatol* **29**(12):2507-12.
- Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, J., Finch, K. L., Stevens, J. F., Livak, K. J., Slotterbeck, B. D., Slifer, S. H., Warren, L. L., Conneally, P. M., Schmechel, D. E., Purvis, I., Pericak-Vance, M. A., Roses, A. D., and Vance, J. M., 2000, SNPing away at complex diseases: analysis of single nucleotide polymorphisms around APOE in Alzheimer disease., *Am. J. Hum. Genet.* **2**:383-94.

- Matteson, E. L., Cohen, M. D., and Conn, D. L., 1998, Rheumatoid arthritis: clinical features and systemic involvement., in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby, London, pp. 5.4.1-5.4.8.
- McIntosh, E., 1996, The cost of rheumatoid arthritis, **35**:781-90
- Meunier, F., Philippi, A., Martinez, M., and Demenais, F., 1997, Affected sib-pair tests for linkage: type I errors with dependent sib- pairs, *Genet Epidemiol* **14**(6):1107-11.
- Mitchel, D., Smith, A., Rowan, B., Warnes, T. W., Haboubi, N. Y., Lucas, S. B., and Chalmers, R. J., 1990, Serum type III procollagen peptide, dynamic liver function tests and hepatic fibrosis in psoriatic patients receiving methotrexate., *Br J Dermatol* **122**(1):1-7.
- Mok, C. C., Lanchbury, J. S., Chan, D. W., and Lau, C. S., 1998, Interleukin-10 promoter polymorphisms in Southern Chinese patients with systemic lupus erythematosus, *Arthritis Rheum* **41**(6):1090-5.
- Mottonen, M., Isomaki, P., Saario, R., Toivanen, P., Punnonen, J., and Lassila, O., 1998, Interleukin-10 inhibits the capacity of synovial macrophages to function as antigen-presenting cells, *Br J Rheumatol* **37**(11):1207-14.
- Mozzato-Chamay, N., Mahdi, O. S., Jallow, O., Mabey, D. C., Bailey, R. L., and Conway, D. J., 2000, Polymorphisms in Candidate Genes and Risk of Scarring Trachoma in a Chlamydia trachomatis-Endemic Population, *J Infect Dis* **182**(5):1545-1548.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H., 1986, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction, *Cold Spring Harb Symp Quant Biol* **51**(Pt 1):263-73.
- Myerscough, A., John, S., Barrett, J. H., Ollier, W. E., and Worthington, J., 2000, Linkage of rheumatoid arthritis to insulin-dependent diabetes mellitus loci: evidence supporting a hypothesis for the existence of common autoimmune susceptibility loci., *Arthritis Rheum* **43**(12):2771-5.
- Nelson, J. L., Hughes, K. A., Smith, A. G., Nisperos, B. B., Branchaud, A. M., and Hansen, J. A., 1993, Maternal-fetal disparity in HLA class II

alloantigens and the pregnancy- induced amelioration of rheumatoid arthritis, *N Engl J Med* **329**(7):466-71.

- Nepom, G. T., and Nepom, B., 1998, Genetics of the major histocompatibility complex in rheumatoid arthritis, in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby International, London, pp. 5.1.1-5.7.12.
- O'Connell, J. R., and Weeks, D. E., 1998, PedCheck: a program for identification of genotype incompatibilities in linkage analysis, *Am J Hum Genet* **63**(1):259-66.
- O'Dell, J. R., Haire, C. E., Erikson, N., Drymalski, W., Palmer, W., Eckhoff, P. J., Garwood, V., Maloley, P., Klassen, L. W., Wees, S., Klein, H., and Moore, G. F., 1996, Treatment of rheumatoid arthritis with methotrexate alone, sulfasalazine and hydroxychloroquine, or a combination of all three medications, *N Engl J Med* **334**(20):1287-91.
- Olerup, O., and Zetterquist, H., 1992, HLA-DR typing by PCR amplification with sequence-specific primers (PCR- SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation, *Tissue Antigens* **39**(5):225-35.
- Ollier, W., and Thomson, W., 1992, Population genetics of rheumatoid arthritis, *Rheum Dis Clin North Am* **18**(4):741-59.
- Otto, J. M., Cs-Szabo, G., Gallagher, J., Velins, S., Mikecz, K., Buzas, E. I., Enders, J. T., Li, Y., Olsen, B. R., and Glant, T. T., 1999, Identification of multiple loci linked to inflammation and autoantibody production by a genome scan of a murine model of rheumatoid arthritis, *Arthritis Rheum* **42**(12):2524-31.
- Parish, L. C., 1963, A historical approach to the nomenclature of rheumatoid arthritis, *Arthritis Rheum* **6**:138-58.
- Perez, H. D., Kimberley, R. P., Kaplan, H. B., Edelson, H., Inman, R. D., and Goldstein, I. M., 1981, Effect of high-dose methylprednisolone

infusion on polymorphonuclear leukocyte function in patients with systemic lupus erythematosus, *Arthritis Rheum* **24**(5):641-7.

- Persellin, R. H., 1976, The effect of pregnancy on rheumatoid arthritis, *Bull Rheum Dis* **27**(9):922-7.
- Phillips, P. E., 1988, Evidence implicating infectious agents in rheumatoid arthritis and juvenile rheumatoid arthritis, *Clin Exp Rheumatol* **6**(1):87-94.
- Pincus, T., Ferraccioli, G., Sokka, T., Larsen, A., Rau, R., Kushner, I., and Wolfe, F., 2002, Evidence from clinical trials and long-term observational studies that disease-modifying anti-rheumatic drugs slow radiographic progression in rheumatoid arthritis: updating a 1983 review., *Rheumatology* **41**(12):1346-56.
- Prockop, D. J., and Kivirikko, K. I., 1995, Collagens: molecular biology, diseases, and potentials for therapy., *Annu Rev Biochem* **64**:403-34.
- Rabinowitz, D., 1997, A transmission disequilibrium test for quantitative trait loci, *Hum Hered* **47**(6):342-50.
- Reveille, J. D., Alarcon, G. S., Fowler, S. E., Pillemer, S. R., Neuner, R., Clegg, D. O., Mikhail, I. S., Trentham, D. E., Leisen, J. C., Bluhm, G., Cooper, S. M., Duncan, H., Tuttleman, M., Heyse, S. P., Sharp, J. T., and Tilley, B., 1996, HLA-DRB1 genes and disease severity in rheumatoid arthritis. The MIRA Trial Group. Minocycline in Rheumatoid Arthritis, *Arthritis Rheum* **39**(11):1802-7.
- Rich, E., Moreland, L. W., and Alarcon, G. S., 1999, Paucity of radiographic progression in rheumatoid arthritis treated with methotrexate as the first disease modifying antirheumatic drug, *J Rheumatol* **26**(2):259-61.
- Risch, N., 1990a, Genetic linkage and complex diseases, with special reference to psychiatric disorders, *Genet Epidemiol* **7**(1):3-16.
- Risch, N., 1990b, Linkage strategies for genetically complex traits. I. Multilocus models, *Am J Hum Genet* **46**(2):222-8.
- Risch, N., 1990c, Linkage strategies for genetically complex traits. II. The power of affected relative pairs, *Am J Hum Genet* **46**(2):229-41.

- Risch, N., Ghosh, S., and Todd, J. A., 1993, Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (NOD) mouse and human insulin-dependent diabetes mellitus (IDDM), *Am J Hum Genet* **53**(3):702-14.
- Robinson, D. R., 1998, Inflammation, in: *Rheumatology* (J. H. Klipple, and P. A. Dieppe, eds.), Mosby, London, pp. 1.7.1-1.7.10.
- Ropes, M. W., Bennett, G. A., Cobb, S., Jacox, R., and Jessar, R. A., 1958, 1958 revision of diagnostic criteria for rheumatoid arthritis, *Bull Rheum Dis* **9**(4):175-6.
- Roses, A. D., 2000, Pharmacogenetics and the practice of medicine., *Nature* **405**(6788):857-65.
- Roses, A. D., 2002a, Pharmacogenetics place in modern medical science and practice., *Life Sci* **70**(13):1471-80.
- Roses, A. D., 2002b, SNPs--where's the beef?, *Pharmacogenomics J.* **2**(5):277-83.
- Rousset, F., Garcia, E., Defrance, T., Peronne, C., Vezzio, N., Hsu, D. H., Kastelein, R., Moore, K. W., and Banchereau, J., 1992, Interleukin 10 is a potent growth and differentiation factor for activated human B lymphocytes, *Proc Natl Acad Sci U S A* **89**(5):1890-3.
- Rudwaleit, M., Tikly, M., Khamashta, M., Gibson, K., Klinke, J., Hughes, G., and Wordsworth, P., 1996, Interethnic differences in the association of tumor necrosis factor promoter polymorphisms with systemic lupus erythematosus, *J Rheumatol* **23**(10):1725-8.
- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A., 1988, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science* **239**(4839):487-91.
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N., 1985, Enzymatic amplification of beta-globin genomic

sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science* **230**(4732):1350-4.

- Salmon, M., Wordsworth, P., Emery, P., Tunn, E., Bacon, P. A., and Bell, J. I., 1993, The association of HLA DR beta alleles with self-limiting and persistent forms of early symmetrical polyarthritis, *Br J Rheumatol* **32**(7):628-30.
- Sangha, O., 2000, Epidemiology of rheumatic diseases, *Rheumatology (Oxford)* **39 Suppl 2**:3-12.
- Sawcer, S., Goodfellow, P. N., and Compston, A., 1997, The genetic analysis of multiple sclerosis., *Trends Genet* **13**(6):234-9.
- Seldin, M. F., Amos, C. I., Ward, R., and Gregersen, P. K., 1999, The genetics revolution and the assault on rheumatoid arthritis, *Arthritis Rheum* **42**(6):1071-9.
- Sham, P., 1998, Statistics in human genetics, in: *Arnold applications of statisctics series* (B. Everitt, ed.), Arnold, London.
- Sham, P. C., and Curtis, D., 1995, An extended transmission/disequilibrium test (TDT) for multi-allele marker loci, *Ann Hum Genet* **59**(Pt 3):323-36.
- Shiozawa, S., Hayashi, S., Tsukamoto, Y., Goko, H., Kawasaki, H., Wada, T., Shimizu, K., Yasuda, N., Kamatani, N., Takasugi, K., Tanaka, Y., Shiozawa, K., and Imura, S., 1998, Identification of the gene loci that predispose to rheumatoid arthritis, *Int Immunol* **10**(12):1891-5.
- Silman, A., Davies, P., Currey, H. L., and Evans, S. J., 1983, Is rheumatoid arthritis becoming less severe?, *J Chronic Dis* **36**(12):891-7.
- Silman, A. J., and Hochberg, M. C., 1993, Epidemiology of the Rheumatic Diseases., Oxford University Press, Oxford.
- Silman, A. J., MacGregor, A. J., Thomson, W., Holligan, S., Carthy, D., Farhan, A., and Ollier, W. E., 1993, Twin concordance rates for rheumatoid arthritis: results from a nationwide study., *Br J Rheumatol* **32**(10):903-7.

- Sobel, E., and Lange, K., 1996, Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics, *Am J Hum Genet* **58**(6):1323-37.
- Sokka, T., Krishnan, E., Hakkinen, A., and Hannonen, P., 2003, Functional disability in rheumatoid arthritis patients compared with a community population in Finland., *Arthritis Rheum* **48**:59-63.
- Solomon, L., Robin, G., and Valkenburg, H. A., 1975, Rheumatoid arthritis in an urban South African Negro population, *Ann Rheum Dis* **34**(2):128-35.
- Spector, T. D., and Hochberg, M. C., 1990, The protective effect of the oral contraceptive pill on rheumatoid arthritis: an overview of the analytic epidemiological studies using meta-analysis, *J Clin Epidemiol* **43**(11):1221-30.
- Spector, T. D., Roman, E., and Silman, A. J., 1990, The pill, parity, and rheumatoid arthritis, *Arthritis Rheum* **33**(6):782-9.
- Spielman, R. S., and Ewens, W. J., 1996, The TDT and other family-based tests for linkage disequilibrium and association, *Am J Hum Genet* **59**(5):983-9.
- Spielman, R. S., McGinnis, R. E., and Ewens, W. J., 1993, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *Am J Hum Genet* **52**(3):506-16.
- Spits, H., and de Waal Malefyt, R., 1992, Functional characterization of human IL-10, *Int Arch Allergy Immunol* **99**(1):8-15.
- Spurdle, A., and Jenkins, T., 1992, Y chromosome probe p49a detects complex PvuII haplotypes and many new TaqI haplotypes in southern African populations, *Am J Hum Genet* **50**(1):107-25.
- Suarez, B. K., and Hampe, C. L., 1994, Linkage and association., *Am J Hum Genet* **54**(3):554-9.
- Suarez, B. K., Hampe, C. L., and van Eerdewegh, P., 1994, Genetic approaches to mental disorders, in: *Genetic approaches to mental*

disorders (E. S. Gershon, and C. R. Cloninger, eds.), American Psychiatric Press, Washington DC, pp. 23 - 46.

- Symmons, D. P., Barrett, E. M., Bankhead, C. R., Scott, D. G., and Silman, A. J., 1994, The incidence of rheumatoid arthritis in the United Kingdom: results from the Norfolk Arthritis Register, *Br J Rheumatol* **33**(8):735-9.
- Symmons, D. P., Turner, G., Webb, R., Asten, P., Barrett, E., Lunt, M., et al, 2002, The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century, Rheumatology, **41**:793-800
- Terwilliger, J. D., and Goring, H. H., 2000, Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design, *Hum Biol* **72**(1):63-132.
- Tindall, K. R., and Kunkel, T. A., 1988, Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase, *Biochemistry* **27**(16):6008-13.
- Tugwell, P., Pincus, T., Yocum, D., Stein, M., Gluck, O., Kraag, G., McKendry, R., Tesser, J., Baker, P., and Wells, G., 1995, Combination therapy with cyclosporine and methotrexate in severe rheumatoid arthritis. The Methotrexate-Cyclosporine Combination Study Group, *N Engl J Med* **333**(3):137-41.
- Turner, D. M., Williams, D. M., Sankaran, D., Lazarus, M., Sinnott, P. J., and Hutchinson, I. V., 1997, An investigation of polymorphism in the interleukin-10 gene promoter, *Eur J Immunogenet* **24**(1):1-8.
- van der Heijde, D. M., van Riel, P. L., Nuver-Zwart, I. H., Gribnau, F. W., and vad de Putte, L. B., 1989, Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis, *Lancet* **1**(8646):1036-8.
- van Montfrans, C., Rodriguez Pena, M., Pronk, I., Ten Kate, F., Te Velde, A., and van Deventer, S., 2002, Prevention of colitis by interleukin 10 transduced T lymphocytes in the SCID mouse transfer model., *Gastroenterology* **123**:1865-76.

- Verhoef, C. M., van Roon, J. A., Vianen, M. E., Lafeber, F. P., and Bijlsma, J. W., 1999, The immune suppressive effect of dexamethasone in rheumatoid arthritis is accompanied by upregulation of interleukin 10 and by differential changes in interferon gamma and interleukin 4 production, *Ann Rheum Dis* **58**(1):49-54.
- Vieira, P., de Waal Malefyt, R., Dang, M.-N., Johnson, K. E., Kastelein, R., Fiorentino, D. F., de Vries, J. E., Roncarolo, M.-G., Mosmann, T. R., and Moore, K. W., 1991, Isolation and expression of human cytokine synthesis inhibitory factor (CSIF/IL-10) cDNA clones:homology to Epstein-Barr virus open reading frame BCRF1., *Proc Natl Acad Sci USA* **88**:1172-1176.
- Vollertsen, R. S., Conn, D. L., Ballard, D. J., Ilstrup, D. M., Kazmar, R. E., and Silverfield, J. C., 1986, Rheumatoid vasculitis: survival and associated risk factors, *Medicine (Baltimore)* **65**(6):365-75.
- Walker, D. J., Griffiths, I. D., and Madeley, D., 1987, Autoantibodies and antibodies to microorganisms in rheumatoid arthritis: comparison of histocompatible siblings, *J Rheumatol* **14**(3):426-8.
- Watkins, H., MacRae, C. A., Fischman, D. A., Seidman, J. G., and Seidman, C. E., 1994, A dinucleotide repeat polymorphism in the MYBPH gene, *Hum Mol Genet* **3**(12):2267.
- Weckmann, A. L., and Alcocer-Varela, J., 1996, Cytokine inhibitors in autoimmune disease, *Semin Arthritis Rheum* **26**(2):539-57.
- Weinblatt, M. E., Keystone, E. C., Furst, D. E., Moreland, L. W., Weisman, M. H., Birbara, C. A., Teoh, L. A., Fischkoff, S. A., and Chartash, E. K., 2003, Adalimumab, a fully human anti-tumor necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial., *Arthritis Rheum* **48**(1):35-45.
- Weinblatt, M. E., Kremer, J. M., Bankhurst, A. D., Bulpitt, K. J., Fleischmann, R. M., Fox, R. I., Jackson, C. G., Lange, M., and Burge, D. J., 1999, A trial of etanercept, a recombinant tumor necrosis factor

receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate., *N Engl J Med* **340**(4):253-9.

- Westendorp, R. G., Langermans, J. A., Huizinga, T. W., Elouali, A. H., Verweij, C. L., Boomsma, D. I., Vandenbroucke, J. P., and Vandenbrouke, J. P., 1997, Genetic influence on cytokine production and fatal meningococcal disease [published erratum appears in Lancet 1997 Mar 1;349(9052):656] [see comments], *Lancet* **349**(9046):170-3.
- Weyand, C. M., and Goronzy, J. J., 1995, Inherited and noninherited risk factors in rheumatoid arthritis, *Curr Opin Rheumatol* **7**(3):206-13.
- Weyand, C. M., Hicok, K. C., Conn, D. L., and Goronzy, J. J., 1992, The influence of HLA-DRB1 genes on disease severity in rheumatoid arthritis, *Ann Intern Med* **117**(10):801-6.
- Wijnands, M. J., and van Riel, P. L., 1995, Management of adverse effects of disease-modifying antirheumatic drugs., *Drug Saf* **13**(4):219-27.
- Wiles, N., Symmons, D., Harrison, B., Barrett, E., Barrett, J. H., Scott, D. G., et al, 1999, Estimating the incidence of rheumatoid arthritis: trying to hit a moving target? **42**:1339-1346
- Wilske, K. R., 1993, Inverting the therapeutic pyramid: observations and recommendations on new directions in rheumatoid arthritis therapy based on the author's experience, *Semin Arthritis Rheum* **23**(2 Suppl 1):11-8.
- Wingrave, S. J., and Kay, C. R., 1978, Royal College of General Practitioners Study. Reduction in incidence of rheumatoid arthritis associated with oral contraceptives., *Lancet* :569-71.
- Wolfe, F., and Cathey, M. A., 1991, The assessment and prediction of functional disability in rheumatoid arthritis [published erratum appears in J Rheumatol 1991 Nov;18(11):1774], *J Rheumatol* **18**(9):1298-306.
- Wolfe, F., Kleinheksel, S. M., Cathey, M. A., Hawley, D. J., Spitz, P. W., and Fries, J. F., 1988, The clinical value of the Stanford Health Assessment Questionnaire Functional Disability Index in patients with rheumatoid arthritis, *J Rheumatol* **15**(10):1480-8.

- Wolfe, F., Mitchell, D. M., Sibley, J. T., Fries, J. F., Bloch, D. A., Williams, C. A., Spitz, P. W., Haga, M., Kleinheksel, S. M., and Cathey, M. A., 1994, The mortality of rheumatoid arthritis, *Arthritis Rheum* **37**(4):481-94.
- Woods, J. M., Tokuhira, M., Berry, J. C., Katschke, K. J., Kurata, H., Damergis, J. A., Arai, K., and Koch, A. E., 1999, Interleukin-4 adenoviral gene therapy reduces production of inflammatory cytokines and prostaglandin E2 by rheumatoid arthritis synovium ex vivo, *J Investig Med* **47**(6):285-92.
- Wordsworth, B. P., 1991, HLA class II antigens in susceptibility to rheumatoid arthritis, *Br J Rheumatol* **30**(2):151-2.
- Wordsworth, B. P., Lanchbury, J. S., Sakkas, L. I., Welsh, K. I., Panayi, G. S., and Bell, J. I., 1989, HLA-DR4 subtype frequencies in rheumatoid arthritis indicate that DRB1 is the major susceptibility locus within the HLA class II region, *Proc Natl Acad Sci U S A* **86**(24):10049-53.
- Wordsworth, P., and Brown, M., 1997, Rheumatoid arthritis and allied inflammatory arthropathies, in: *Principles and practice of medical genetics* (D. L. Rimoin, J. M. Connor, and R. E. Pyeritz, eds.), Churchill Livingstone, pp. 2751 - 2771.
- Wordsworth, P., and Brown, M., 1998, Rheumatoid arthritis and allied inflammatory arthropathies., in: *Principles and practice of medical genetics.* (D. L. Rimoin, J. M. Connor, and R. E. Pyeritz, eds.), pp. 2751-2771.
- Yelin, E., Henke, C., and Epstein, W., 1987, The work dynamics of the person with rheumatoid arthritis, *Arthritis Rheum* **30**(5::507-12..

Appendices

5.1.1 Core set of variables to assess disease activity in rheumatoid arthritis

The table below lists a set of core variables designed to assess disease activity in RA. There are a number of instruments designed to measure this and most use the same core variables. The three most commonly used measures are listed below.

Disease activity measure	Core sets		
	ACR [$]	EULAR *	OMERACT ^
Tender joint count	√ (68) [§]	√ (28)[§]	?
Swollen joint count	√ (68) [§]	√ (28) [§]	?
Pain on VAS*	√	√	√
Patients global assessment of disease activity	√	√	√
Assessors global assessment of disease activity	√		√
Physical function	√	√	√
Radiographic analysis	√	√	√
Acute-phase reactant	√	√	√

Adapted from van Reil PLCM, van de Putte LBA [van Reil, 1994 #1219]

VAS = visual analogue scale

[§] numbers in brackets indicate the number of joints assessed

* EULAR = European League against Rheumatism

[$] ACR = American College of Rheumatology

^ OMERACT = Outcome Measures in RA Clinical Trials

5.1.2 EULAR response definition

Unlike the Paulus [Paulus, 1990 #1095] and ACR [Felson, 1995; Felson, 1993] response criteria (appendix 5.1.4) the EULAR response criteria [van Gestel, 1996] are based on a combination of current disease activity *and* change in disease activity.

Current DAS [§]	Decrease in DAS		
	> 1.2	> 0.6 and ≤ 1.2	≤ 0.6
≤ 2.4	GOOD	MODERATE	NONE
> 2.4 and ≤ 3.7	MODERATE	MODERATE	NONE
> 3.7	MODERATE	NONE	NONE

EULAR = European League against Rheumatism

[§] DAS = disease activity score

ACR criteria for the clinical remission of Rheumatoid Arthritis

Table 5.1.3a

Criterion	Description
1	Early morning stiffness not exceeding 15 minutes
2	No fatigue
3	No joint pain
4	No joint tenderness or pain on motion
5	No soft tissue swelling in joints or tendon sheaths
6	ESR < 30mm/h in females or < 20mm/h in males

Table 5.1.3b

Exclusions	Diagnosis
1	Clinical manifestations of active vasculitis
2	Pericarditis
3	Pleuritis
4	Myositis
5	+/or unexplained weight loss or fever secondary to rheumatoid arthritis

Tables 5.1.3a and b have been adapted from Gordon [Gordon, 1998 #704]. Table 5.1.3a lists the criteria used to determine clinical remission of RA. A minimum of five out of the six criteria listed in the table need to be fulfilled for a minimum of two months before remission can be diagnosed. However there are some exclusions and these are listed in table 5.1.3b. Hence an individual cannot be described as being in remission if they fulfil the criteria listed in table 1.1.4a but have the clinical manifestations of active vasculitis or a pericarditis.

The American College of Rheumatology criteria for classification of global functional status in rheumatoid arthritis

The American College of Rheumatology criteria for classification of global functional status in rheumatoid arthritis	
Class I	Completely able to perform usual activities of daily living (self-care, vocational and avocational)
Class II	Able to perform usual self-care and vocational activities, but limited in avocational activities
Class III	Able to perform usual self-care activities but limited in vocational and avocational activities
Class IV	Limited in ability to perform usual self-care, vocational and avocational activities

Table 5.1.5 summarises the American College of Rheumatology criteria for the classification of global functional status in rheumatoid arthritis [Hochberg, 1992]. The criteria relate to an individual's ability to self-care, work and undertake any activities or hobbies they may wish to. The higher the class, the poorer the function and the more severe the disease. One of the aims of treatment is to maintain an individual in the lowest class possible.

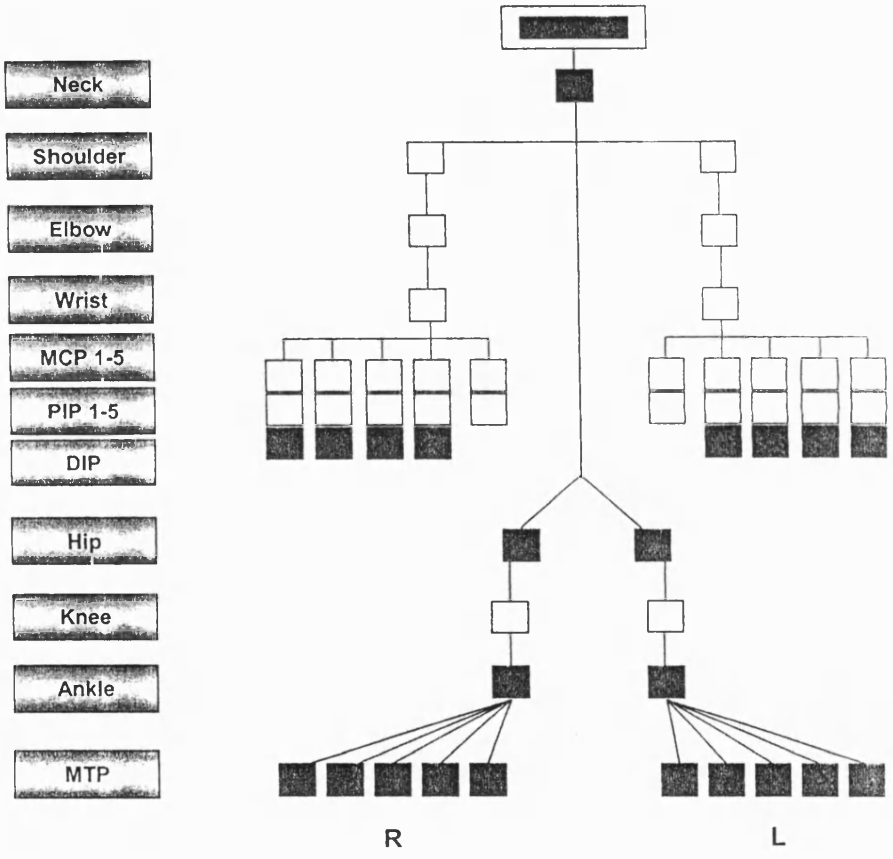
Date: d d m m y y y y

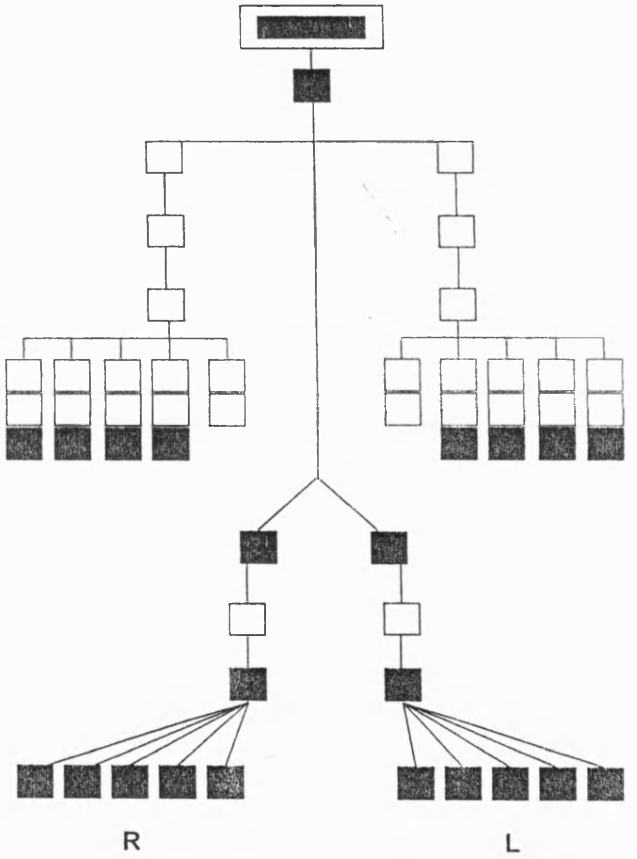
28-JOINT SWOLLEN AND TENDER JOINT COUNT

Patient Name _____

Which joints are **tender**? (please tick)

Which joints are **swollen**? (please tick)





Global VAS: Overall well-being: Please indicate on the scale below

0 ---------- 100

Best Imaginable Health State — Worst Imaginable Health State

_____

ESR

CRP

You do not need to return this sheet: it is to help you complete the questionnaire

256

HEALTH ASSESSMENT QUESTIONNAIRE (HAQ)

Date: | | | | | | | | | **Patient Name:**

Please tick the one response which best describes your usual abilities over the past week

	Without ANY difficulty	With SOME difficulty	With MUCH difficulty	UNABLE to do	
1. DRESSING and GROOMING					
Are you able to:					
a. Dress yourself, including tying shoelaces and doing buttons?	☐	☐	☐	☐	---
b. Shampoo your hair?	☐	☐	☐	☐	
2. RISING					
Are you able to:					
a. Stand up from an armless straight chair?	☐	☐	☐	☐	---
b. Get in and out of bed?	☐	☐	☐	☐	
3. EATING					
Are you able to:					
a. Cut your meat?	☐	☐	☐	☐	---
b. Lift a full cup or glass to your mouth?	☐	☐	☐	☐	
c. Open a new carton of milk (or soap powder)?	☐	☐	☐	☐	
4. WALKING					
Are you able to:					
a. Walk outdoors on flat ground?	☐	☐	☐	☐	---
b. Climb up five steps?	☐	☐	☐	☐	

PLEASE TICK ANY AIDS OR DEVICES THAT YOU USUALLY USE FOR ANY OF THESE ACTIVITIES:

Cane (W)	☐	Walking frame(W)	☐	Built-up or special utensils (E)	☐
Crutches (W)	☐	Wheelchair (W)	☐	Special or built-up chair (A)	☐

Devices used for dressing (button hooks, zipper pull, shoe horn) ☐

Other (specify).....

PLEASE TICK ANY CATEGORIES FOR WHICH YOU USUALLY NEED HELP FROM ANOTHER PERSON:

Dressing and Grooming	☐	Eating	☐
Rising	☐	Walking	☐

ID | | | | | | | | |

For office use only

Please tick the one response which best describes your usual abilities over the past week

	Without ANY difficulty	With SOME difficulty	With MUCH difficulty	UNABLE to do
5. HYGIENE				
Are you able to:				
a. Wash and dry your entire body?	☐	☐	☐	☐
b. Take a bath?	☐	☐	☐	☐
c. Get on and off the toilet?	☐	☐	☐	☐
6. REACH				
Are you able to:				
a. Reach and get down a 5 lb object (e.g. a bag of potatoes) from just above your head?	☐	☐	☐	☐
b. Bend down to pick up clothing off the floor?	☐	☐	☐	☐
7. GRIP				
Are you able to:				
a. Open car doors?	☐	☐	☐	☐
b. Open jars which have been previously opened?	☐	☐	☐	☐
c. Turn taps on and off?	☐	☐	☐	☐
8. ACTIVITIES				
Are you able to:				
a. Run errands and shop?	☐	☐	☐	☐
b. Get in and out of a car?	☐	☐	☐	☐
c. Do chores such as vacuuming, housework or light gardening?	☐	☐	☐	☐

PLEASE TICK ANY AIDS OR DEVICES THAT YOU USUALLY USE FOR ANY OF THESE ACTIVITIES:

Raised toilet seat (H) ☐ Bath seat (H) ☐ Bath rail (H) ☐

Long handled appliances for reach (R) ☐

Jar opener (for jars previously opened) (G) ☐

Other (specify) _____

PLEASE TICK ANY CATEGORIES FOR WHICH YOU USUALLY NEED HELP FROM ANOTHER PERSON:

Hygiene ☐ Gripping and opening things ☐

Reach ☐ Errands and housework ☐

HAQ
0 0.000
1 0.125
2 0.250
3 0.375
4 0.500
5 0.625
6 0.750
7 0.875
8 1.000
9 1.125
10 1.250
11 1.375
12 1.500
13 1.625
14 1.750
15 1.875
16 2.000
17 2.125
18 2.250
19 2.375
20 2.500
21 2.625
22 2.750
23 2.875
24 3.000

ID ☐☐☐☐☐☐☐☐

For office use only

Patient Name: _____

SF-36 v2

Today's Date

c	c	m	m	y	y	y	y

Patient ID

For office use

The following questions ask for your views about your health, how you feel and how well you are able to do your usual activities. If you are unsure about how to answer any questions please give the best answer you can. Do not spend too much time in answering, as your immediate answer is likely to be the most accurate.

---

1. **In general**, would you say your health is: **(Please tick one)**

Excellent	Very Good	Good	Fair	Poor
○	○	○	○	○

---

2. **Compared to one year ago**, how would you rate your health in general now? **(Please tick one)**

Much better now than one year ago	Somewhat better now than one year ago	About the same as one year ago	Somewhat worse now than one year ago	Much worse now than one year ago
○	○	○	○	○

---

3. The following questions are about activities you might do during a typical day **(Please tick one)**

Does your health now limit you in your activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
a) **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports?	○	○	○
b) **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf?	○	○	○
c) Lifting or carrying groceries?	○	○	○
d) Climbing **several** flights of stairs?	○	○	○
e) Climbing **one** flight of stairs?	○	○	○
f) Bending, kneeling or stooping?	○	○	○
g) Walking **more than one mile**?	○	○	○
h) Walking **several hundred yards**?	○	○	○
i) Walking **100 yards**?	○	○	○
j) Bathing or dressing yourself?	○	○	○

Patient ID

4. During the **past four weeks**, how much of the time have you had any of the following problems with your work or other regular daily activities as **a result of your physical health**?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Cut down on the **amount of time** you spent on work or other activities?	○	○	○	○	○
b) **Accomplished** less than you would like?	○	○	○	○	○
c) Were limited in the **kind** of work or other activities?	○	○	○	○	○
d) Had **difficulty** performing the work or other activities, i.e. it took extra effort?	○	○	○	○	○

---

5. During the **past four weeks**, how much of the time have you had any of the following problems with your work or other regular daily activities as **a result of any emotional problems** (such as feeling depressed or anxious)? **Please tick one**

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Cut down on the **amount of time** you spent on work or other activities?	○	○	○	○	○
b) **Accomplished less** than you would like to?	○	○	○	○	○
c) Did work or other activities **less carefully than usual**?	○	○	○	○	○

---

6. During the **past four weeks**, to what extent has your **physical health or emotional problems** interfered with your normal social activities with family, friends, neighbours or groups? **Please tick one**

Not at all	Slighty	Moderately	Quite a bit	Extremely
○	○	○	○	○

---

7. How much **bodily** pain have you had during **the past four weeks?** **Please tick one**

None	Very mild	Mild	Moderate	Severe	Very severe
○	○	○	○	○	○

---

8. During the **past four weeks**, how much did **pain** interfere with your normal work? **Please tick one**

(including work both outside the home and housework)

Not at all	A little bit	Moderately	Quite a bit	Extremely
○	○	○	○	○

Patient ID

9. The following questions are about how you feel and how things have been with you **during the past 4 weeks**. (for each question, please indicate the one answer that comes closest to the way you have been feeling)

How much time during the past 4 weeks:

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Did you feel full of life?	○	○	○	○	○
b) Have you been very nervous?	○	○	○	○	○
c) Have you felt so down in the dumps that nothing could cheer you up?	○	○	○	○	○
d) Have you felt calm and peaceful?	○	○	○	○	○
e) Did you have a lot of energy?	○	○	○	○	○
f) Have you felt downhearted and depressed?	○	○	○	○	○
g) Did you feel worn out?	○	○	○	○	○
h) Have you been happy?	○	○	○	○	○
i) Did you feel tired?	○	○	○	○	○

10. During the **past four weeks**, how much of the time has your **physical health or emotional problems** interfered with your social activities? (like visiting friends, relatives, etc...) **Please tick one**

All of the time	Most of the time	Some of the time	A little of the time	None of the time
○	○	○	○	○

11. How **TRUE** or **FALSE** are each of the following statements? **Please tick one**

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
a) I seem to get ill more easily than other people	○	○	○	○	○
b) I am as healthy as anybody I know	○	○	○	○	○
c) I expect my health to get worse	○	○	○	○	○
d) My health is excellent	○	○	○	○	○

THANK YOU FOR TAKING THE TIME TO COMPLETE THIS QUESTIONNAIRE.

5.2 Definitions

Analysis – Parametric linkage analysis

Linkage analysis requiring a model for the segregation of the trait and marker data in multigenerational pedigrees. This analysis is quite sensitive to model mis-specification.

Analysis – Non-parametric linkage analysis

Used for complex traits where the *a priori* model of trait segregation is unknown. This form of analysis only requires the marker and trait data but no model for segregation.

Association

A population association between two traits exists if the population frequency of persons with both traits is not equal to the product of the population frequency of persons with each individual trait, e.g. blood group and disease. Association of a marker locus with a disease locus is a population-based phenomenon.

Allelic association

Allelic association refers to a significantly increased or decreased frequency of occurrence of a marker allele in combination with a disease trait. It can be explained either by biological interaction of the marker allele with the disease-trait gene or by linkage disequilibrium.

Bonferroni correction

It is inherent in the definition of p-value and significance levels that some false positive results will be generated. The Bonferroni correction assumes that the tests that have been applied to a data set are mutually independent and sets the threshold of significance at $p=0.05/n$, where n is the number of independent potential associations checked. If n loci with m alleles each are tested the rigorous correction factor would be $n(m-1)$.

Causality

Any association identified could be due to direct causation, chance, bias, or confounding factors. Causality is more likely if the exposure precedes outcome and the association is biologically plausible, strong, consistent between studies or supported by good evidence.

Classification criteria

The standards of classification criteria include validity or the accurate definition of disease and an appropriate selection of cases and controls. The potential criteria selected should be comprehensive and there should be absence of circular reasoning (i.e. avoiding the identification of criteria identical to those used to select cases). The criteria need to be validated independently and any statistical techniques need to be appropriately applied.

Epidemiology – analytical

Analytical epidemiology tests hypotheses of causation and involves a comparison group e.g. a prospective or retrospective case-control or cohort study.

Epidemiology – descriptive

Descriptive epidemiology investigates the distribution of disease in relation to person, place and time.

Epistasis

Two or more loci show epistasis if they interact i.e. if the alleles at each locus influence the phenotype produced by the other locus.

Errors

A type I error (α) is a false positive and the null hypothesis is rejected when it is true. The correct decision is $1-\alpha$.

A type II error (β) is a false negative and the null hypothesis is not rejected when it is false. The correct decision is $1-\beta$.

Genotype

The frequency of each genotype in a population is a function of the frequencies of the alleles in that population. For a gene with *n* alleles there are n(n+1)/2 possible genotypes.

Heterogeneity - allelic or intragenic

This occurs when many different mutations within a given gene product produce the disease in different families e.g. cystic fibrosis, thalassemia or X linked hypophosphataemic rickets. Allelic heterogeneity cannot be detected by linkage analysis because all mutations occur within the same locus. However, different mutations within the same gene also can cause different diseases e.g. different mutations of type 1 collegen can cause different types of osteogenesis imperfecta or Ehlers Danlos syndrome.

Heterogeneity - genetic

Different mutations of a genotype cause similar phenotypes.

Heterogeneity - locus

Locus heterogeneity is expected in conditions where a general pathway has failed e.g. blindness, mental retardation or deafness. It covers the situation where defects in different genes produce the same clinical abnormality. For instance profound hearing loss can occur as a result of defects in number of different genes and is an example of complementation. This can often be detected by linkage analysis as the defects are at different loci.

Heterozygosity

Heterozygosity is a measure of the informativeness of a marker locus. It represents the extent of genetic variation of a marker locus within a given population i.e. the frequency of heterozygote individuals.

Incidence

The number of individuals within a defined population who first develop the disorder over a given time period e.g. one year.

Likelihood

A likelihood (L) is the numerical estimate of the fit of the pedigree data to a particular genetic transmission model.

Linkage analysis

The primary goal is to determine whether two or more genetic traits (the marker locus (or multiple marker loci) and the disease trait) are co-segregating within a pedigree.

Linkage disequilibrium

Linkage disequilibrium occurs when a particular marker allele is associated with the disease-trait locus at a greater than expected frequency across multiple families.

LOD score

The LOD score was defined by Morton in 1955 and represents $\log_{10}$ of the odds for linkage. LOD is a contraction of 'logarithm of the odds ratio'. The two hypotheses being compared in linkage analysis are whether the trait is linked or unlinked to the marker (i.e. the recombination fraction between the trait and the marker is $\theta < 0.5$ (linkage) or $\theta = 0.5$ (no linkage). Conventionally a LOD score of 3.0 is regarded as the minimum criteria for claiming linkage and this corresponds to odds of 1000 to 1 in favour of linkage. A LOD score of –2.0 is the conventional threshold for excluding a site from linkage.

Maximum likelihood method

This was described by Morton in 1959. This technique permits joint consideration of several parameters of the genetic model including: degree of dominance, penetrance, gene frequency, ascertainment probability and measures of Mendelian transmission of the disease allele. To find the most likely hypothesis, parameter values are set to maximise the likelihood of the observed data. The parameter values are called the maximum likelihood estimates. This method can be used to establish whether or not the disease has a major genetic component.

Models for genetic interaction

A multifactorial disease will generally involve two or more genetic loci that may interact in several ways (epistasis). In an additive model the penetrance

of a disease is represented by the *sum* of the separate terms contributed by two or more loci. In a multiplicative model, the penetrance of a disease is represented by the *product* of separate terms (or penetrance factors) contributed by two or more loci. The heterogeneity model assumes that genetic loci act independently (i.e. not epistatically) so that an individual can become affected through possessing a particular genotype at a locus regardless of the alleles at the other loci.

Mortality

$$\frac{\text{number of deaths in population studied}}{\text{total population}}$$

Mortality (Case-fatality)

$$\frac{\text{number of deaths}}{\text{number of cases}}$$

Multi-factorial inheritance

Multi-factorial inheritance is an extension of polygenic inheritance where additional nongenetic factors may also be involved.

Mutation

This is a change in DNA sequence, usually conferring a deleterious effect.

Odds ratio and Relative risk

	Disease (+)	Disease (-)	
Exposure (+)	a	b	a+b
Exposure (-)	c	d	c+d
	a+c	b+d	

$$\text{Odds ratio (OR)} = \frac{\text{odds of disease in exposed cohort}}{\text{Odds of disease in unexposed cohort}} = \frac{ad}{cb}$$

$$\text{Relative risk (RR)} = \frac{\text{risk (incidence) of disease in exposed cohort}}{\text{risk in unexposed cohort}} = \frac{a(c+d)}{c(a+b)}$$

Oligenic inheritance

A trait is oligogenic if a few genetic loci acting together cause the trait or disease.

Penetrance

The penetrance of a particular genotype at a disease locus is the probability that individuals possessing that genotype will develop that disease. The penetrance of a disease is complete if *all* individuals with the high-risk genotype will eventually develop the disorder (if they live for long enough). The penetrance of a disease is incomplete, reduced or partial if not all individuals possessing the disease genotype develop the trait. Incomplete pentrances may be a function of many variables including age of onset, sex, environment and the presence of certain alleles at other genetic loci. If the mutant gene is expressed subclinically the disease may appear non-penetrant upon clinical examination e.g. Wilson disease.

Phenotype

A phenotype is an observable biochemical, cellular or clinical characteristic that is associated with the expression of a gene.

Pleiotropy

This is where the risk association can be attributed to a *specific* allele. This is different from linkage disequilibrium where the risk is associated with an *unknown* allele.

Polygenic inheritance

In mathematical terms, a polygenic trait is one in which there are an infinite number of genes each exerting a tiny and equal influence on the trait value. However, the term generally implies that the disease is the result of the additive effects of many of different genes, each of which, acting alone, would be insufficient to cause the disease. Each of the contributory genes may not have equal weight; some may have major effects in conferring risk for the disease, whereas others might provide a relatively minor contribution.

Polymorphism

DNA sequence difference, usually of no pathologic significance. For a locus to be polymorphic there must be identifiable alternative forms (alleles) at that locus. A locus is generally considered to be polymorphic if it has at least two alleles each with a population frequency ≥ 0.01. Any allele with a frequency ≥ 0.01 is termed a variant

Population attributable risk fraction

This is the proportion of the disease in the population attributable to the risk factor e.g. not all GI bleeds can be attributed to NSAIDs

Power

The power of a test is the probability of rejecting the null hypothesis given an alternative hypothesis is true. It is defined in the context of specific circumstances.

Prevalence - point prevalence

The proportion of the population with the disorder at a particular time

Prevalence - period prevalence

The proportion of the population with the disorder at some stage during a given time period e.g. one year

Prevalence - cumulative prevalence

The proportion of the population who have had the disorder by a specified time point e.g. age 60 years

Probability

Events are *independent* if knowledge of any subset of them does not allow any inferences to be made about the remainder. However, the disease prevalence for a sibling of an affected individual is a *conditional probability* i.e. it depends upon the sibling recurrence risk λ_s. The *prior* probability of some event is the probability assigned to that event before any new tests have been carried out. The *posterior* probability is the probability calculated using the prior probability and any new information gathered.

Qualitative trait

An example of a qualitative trait is one where the individual either has the trait or does not e.g. eye colour.

Quantitative trait

Quantitative traits are phenotypes that can fall anywhere along a continuous distribution of measurements e.g. height

Segregation analysis

The primary goal is to determine the most likely genetic model for an observed trait – i.e. to delineate whether the trait is autosomal dominant, autosomal recessive, X-linked or polygenic. To investigate mode of

inheritance by segregation analysis accurate phenotype and genealogical data on a set of families must be obtained. The means by which families are ascertained must be uniform to avoid bias and incidence of the disease in the general population as well as sex and age dependencies must be incorporated in the analysis. Common diseases or those associated with reduced penetrance will necessitate a larger sampling of families.

Single-gene inheritance

Any trait caused by a single gene mutation is said to be Mendelian or to have single gene inheritance e.g. Huntington disease or cystic fibrosis.

Variable expressivity

Expressivity of a genetically transmitted disease trait refers to the severity of the phenotype.

DNA EXTRACTION PROTOCOL

ISOLATION OF DNA FROM BLOOD USING GUANIDINE HYDROCHLORIDE

- The following protocol works best on whole blood which has been frozen at -20C for **at least** 24hrs, although it does work for fresh blood.
- Digestion of WBC pellet in step 6 can be done at 60C for 2hrs or at 37C overnight.
- Set water-bath at chosen temp, before starting procedure, so ready when needed.

STAGE 1: RBC LYSIS

1. Defrost blood thoroughly at room temp (2-3hrs). The centrifuge capacity is frequently 16 tubes, hence this is the recommended number of DNA extractions per run.
2. In fume hood, pour 10-20ml of blood into 50ml Falcon tube and make up to 40ml with **cold** lysis buffer and vortex.
3. Centrifuge at 2,500rpm for 15 minutes then gently tip off the supernatant into waste pot being careful not to disturb the pellet
4. Add 20ml of **cold** lysis buffer to pellet and vortex until the pellet is completely dispersed. *Note:- this step is very important! Ensure that the pellet is fully dispersed.*
5. Centrifuge at 2,500rpm for 15 minutes then gently tip off the supernatant into waste pot to leave clean WBC pellet.
6. If the blood is old and/or clotted, it may be necessary to redo the washing steps (steps 4-5 above) again until a pinkish-white pellet is obtained.

STAGE 2: GUANIDINE-HCL DNA EXTRACTION

1. Add 3.5 mls of 6M guanidine hydrochloride (GuHCl) and vortex to re-suspend pellet.

2. Add 250mls of 7.5M ammonium acetate (NH4Ac) and mix by inversion.
3. Add 50μl of Proteinase K at 10mg/ml then add 250ul of 10% sodium dodecyl sulphate (SDS) and shake. (*Note the SDS tends to precipitate out as lumps.*)
4. Incubate at 37°C overnight or at 60°C for 1 hour.
5. Cool tube to room temperature and add 2ml COLD Chloroform ($CHCl_3$) - vortex well.
6. Centrifuge at 3000 rpm for 5 minutes.
7. Use disposable plastic pipette to collect the upper layer into a clean 50ml Falcon tube and add 10mls of 100% cold ethanol. Mix thoroughly by inversion.
8. Centrifuge briefly at 3000 rpm, drain supernatant and add 10ml 80% ethanol (*use 80% Ethanol only, otherwise DNA begins to re-suspend*), gently swirl to dislodge pellet.
9. Repeat steps 7 and 8.
10. Drain the ethanol off and re-spin to collect the last drops of ethanol at the bottom of the tube. Remove the last traces of ethanol with a yellow tip.
11. Leave the Falcon tube open for 15-20 minutes or until the DNA pellet is visibly dry.
12. Add 1ml of 1xTE (pH 8.0), replace the lid and leave overnight at room temperature to re-suspend.
13. Transfer to eppendorf tubes, ensuring that the DNA is fully re-suspended (clean DNA which has not been over-dried will go into solution very easily). If it does not re-suspend easily warm to 37°C for 1-2 hours.

RE-EXTRACTION OF SAMPLES

1. Split the samples into eppendorf tubes (up to 500µl of DNA per tube).
2. Add 500µl of a 1:1 mixture of phenol and Chloroform ($CHCl_3$).
3. Vigorously mix by inversion until an emulsion is formed.
4. Centrifuge at top speed in the microfuge for 5 minutes.
5. Using a blue tip carefully transfer the top layer into a clean eppendorf (do not transfer the interphase - this contains the contaminating protein).
6. Add 1ml of 100% Ethanol and mix by inversion.
7. Pellet the DNA by briefly spinning in the microfuge at top speed.
8. Remove all of the Ethanol.
9. Add 800µl of 80% Ethanol, ensuring the pellet is dislodged. If the pellet sticks then flick the tube until it is washed in the Ethanol.
10. Centrifuge briefly to collect the DNA in the bottom.
11. Remove all of the 80% Ethanol. Re-centrifuge and remove the last traces of 80% Ethanol with a yellow tip.
12. Leave the samples to stand open for 10-15 minutes or until visibly dry.
13. Add 400µl of 1xTE (pH 8.0) and allow to re-dissolve.
14. Re-analyse by spectrophotometry as above.

TO MAKE LYSIS BUFFER

1. Place 109.4g of Sucrose in a 1L Schott bottle.
2. Add ~500 mls of distilled water, 1 ml of 1M (molar) Tris-HCl* (pH 7.8), 1.2 mls 4.9M $MgCl_2$ § and 10 mls of Triton ×-100.
3. Mix by swirling until the sucrose has more or less dissolved.
4. Make up to 1 L with Milli-Q water. The sucrose should dissolve after a while - store at 4°C.

* Tris = hydroxy-methyl-amino-methane

§ $MgCl_2$ = magnesium chloride

TO MAKE 6M GUANIDINIUM HYDROCHLORIDE (GuHCl)

To make up 500 mls:

1. Dissolve 286.59g of solid guanidine hydrochloride in ~300 mls of distilled (Milli-Q) water.
2. Make up to 500 mls with Milli-Q water.
3. Sterilize by autoclaving.

TO MAKE 7.5M AMMONIUM ACETATE (NH_4Ac)

To make up 100 mls:

1. Dissolve 57.81g of solid NH_4 Ac. in ~75 mls of Milli-Q water.
2. Make up to 100 mls with Milli-Q water.
3. Sterilize by autoclaving.

TO MAKE 10% SDS (SODIUM DODECYL SULPHATE)

To make up 100 mls:

1. Dissolve 10g SDS powder in ~75 mls Milli-Q water.
2. Make up to 100 mls with Milli-Q water.

Tris-EDTA BUFFER (TE8)

1. take 1ml of Tris-EDTA buffer (1M Tris 100* concentrate)
2. dilute in 99ml distilled water

SPECTROPHOTOMETRIC ASSESSMENT

1. Take 2µl of the extracted DNA sample and add to 98µl of sterile de-ionised water (SDW), vortex.
2. Switch on the UV lamp on the spectrophotometer (leave to warm up).
3. Quit any windows currently open on the spectrophotometer, then open 'DNA analysis'.
4. Put 100µl of SDW into a cuvette and place in the spectrophotometer.
5. Press 'Read Blank'.
6. Wash out the cuvette and replace with the sample.
7. Press 'Read Samples'.
8. Record the OD_{260} and the ratio of OD_{260}/OD_{280}. If the ratio is 1.8 - 2.0 then the DNA is very clean but if the ratio is less than 1.5, then the sample is heavily contaminated with protein and needs to be re-extracted.
9. Repeat steps 6-8, but redo the blank for every 8th sample.

EQUATION:

- 1 optical density @ 260nm = 50ng/µl dsDNA
- concentration of DNA [ng/µl] = optical density × 50 × dilution factor
- concentration of DNA [ng/µl] = optical density × 50 × 50 (i.e. 2500)

DNA is often diluted to 10ng/µl (in 500µl water) therefore to work out quantity of DNA stock required:

10 × 500 / concentration of DNA = µl DNA needed

(hence water needed = 500µl - µl DNA needed)

i.e. for any concentration of DNA in any volume =

ng needed × volume needed / concentration of DNA [ng/µl]

"HYDRA" GUIDELINES

- MAIN SWITCH is at the back of the *HYDRA*, at the power connection point.
- SET/RESET button takes you through the different files and through the steps of each file.
- Buttons FILL, EMPTY and WASH only function at the end of each file, when the display is showing 'A' (for aspirate) or 'D' (for dispense):

  FILL - fills the syringes up to the maximum (290µl)

  EMPTY - empties the syringes

  WASH - starts the wash cycle (max. fill and empty, 3 times)

Files 1-6 have been programmed as follows:

FILE 1	Aspirating from clear boxes
FILE 2	Dispensing into clear boxes (to get rid of bubbles)
FILE 3	Dispensing into U-bottom plates
FILE 4	Dispensing into V-bottom plates
FILE 5	Aspirating from V-bottom plates
FILE 6	Aspirating from 'Beckman' boxes

These files alter the height of the platform only while aspirating/dispensing liquid. The volume to be aspirated/dispensed can be varied and can be set by pressing the UP or DOWN buttons when the display is showing AV (aspiration volume) or DV (dispensing volume).

The heights of the platform are set up by pressing the UP button, with the chosen box/plate in place, while watching closely until the needles are about 3mm from the bottom of the tubes/wells.

OPERATING THE *HYDRA*

- Place the box/plate (containing the liquid to be transferred) onto the platform.

- Press SET/RESET button until the display shows 'File'. Pressing UP button will change the file number.
- Once in the correct file, press SET/RESET button until the display shows AV or DV. This also indicates whether the set function is aspiration or dispensing (pressing the UP button will change between the two modes).
- Adjust the desired volume to be aspirated or dispensed.
- Keep pressing the SET/RESET button to go through all the steps of the file.
- When the display shows A or D, press the ASPIRATE/DISPENSE button which will perform the set operation.

ARTHRITIS & RHEUMATISM
Vol. 46, No. 3, March 2002, pp 632–639
DOI 10.1002/art.10147

Whole-Genome Linkage Analysis of Rheumatoid Arthritis Susceptibility Loci in 252 Affected Sibling Pairs in the United Kingdom

Kirsten MacKay,[1] Stephen Eyre,[2] Anne Myerscough,[2] Anita Milicic,[1] Anne Barton,[2] Steven Laval,[1] Jenny Barrett,[2] Dorothea Lee,[1] Sarah White,[2] Sally John,[2] Matthew A. Brown,[1] John Bell,[1] Alan Silman,[2] William Ollier,[2] Paul Wordsworth,[1] and Jane Worthington[2]

Objective. **To undertake a systematic whole-genome screen to identify regions exhibiting genetic linkage to rheumatoid arthritis (RA).**

Methods. **Two hundred fifty-two RA-affected sibling pairs from 182 UK families were genotyped using 365 highly informative microsatellite markers. Microsatellite genotyping was performed using fluorescent polymerase chain reaction primers and semiautomated DNA sequencing technology. Linkage analysis was undertaken using MAPMAKER/SIBS for single-point and multipoint analysis.**

Results. **Significant linkage (maximum logarithm of odds score 4.7 [$P = 0.000003$] at marker *D6S276*, 1 cM from HLA–DRB1) was identified around the major histocompatibility complex (MHC) region on chromosome 6. Suggestive linkage ($P < 7.4 \times 10^{-4}$) was identified on chromosome 6q by single- and multipoint analysis. Ten other sites of nominal linkage ($P < 0.05$) were identified on chromosomes 3p, 4q, 7p, 2 regions of 10q, 2 regions of 14q, 16p, 21q, and Xq by single-point analysis and on 3 sites (1q, 14q, and 14q) by multipoint analysis.**

Conclusion. **Linkage to the MHC region was confirmed. Eleven non-HLA regions demonstrated evidence of suggestive or nominal linkage, but none reached the genome-wide threshold for significant linkage ($P = 2.2 \times 10^{-5}$). Results of previous genome screens have suggested that 6 of these regions may be involved in RA susceptibility.**

Rheumatoid arthritis (RA) is a chronic, systemic, inflammatory condition predominantly involving synovial joints and affecting up to 1% of the population. Evidence from twin (1,2) and family (3) studies suggests that both genetic and environmental factors contribute to susceptibility to RA, and disease heritability has been estimated to be ~60% (4). To date, only the major histocompatibility complex (MHC) on chromosome 6 has been consistently linked to and associated with RA susceptibility (5). However, results of family studies suggest that this association accounts for only one-third of the genetic susceptibility (6,7). It is likely that a number of other susceptibility loci exist, each of which contributes less than the MHC to the total genetic component.

If appropriately powered, systematic linkage screening of the entire genome has the potential to detect all disease susceptibility loci. Results from 3 previous whole-genome scans in Japanese (8), European (European Consortium on Rheumatoid Arthritis Families [ECRAF]) (9), and North American (North American Rheumatoid Arthritis Consortium [NARAC]) (10) RA families have been reported, and linkage to the MHC ($P < 2.5 \times 10^{-5}$) was confirmed in the European and North American studies. In all 3 studies, a number of non-MHC regions with evidence of nominal linkage ($P < 0.05$) were also identified. It is inevitable that many suspected linkages will not be confirmed; therefore, it is

Ms. Milicic's work was supported by the Grenville Bequest for Research into Genetic Susceptibility to Rheumatoid Arthritis.

[1]Kirsten MacKay, MRCP, Anita Milicic, MSc, Steven Laval, PhD, Dorothea Lee, BSc, Matthew A. Brown, MD, John Bell, MD, Paul Wordsworth, FRCP: Wellcome Trust Centre for Human Genetics, Oxford, UK; [2]Stephen Eyre, MSc, Anne Myerscough, BSc, Anne Barton, MRCP, Jenny Barrett, PhD, Sarah White, BSc, Sally John, PhD, Alan Silman, FRCP, William Ollier, PhD, Jane Worthington, PhD: University of Manchester, Manchester, UK.

Address correspondence and reprint requests to Jane Worthington, PhD, ARC Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester M13 9PT, UK. E-mail: Jane@fs1.ser.man.ac.uk.

Submitted for publication June 12, 2001; accepted in revised form November 2, 2001.

Table 1. Summary of potential rheumatoid arthritis loci identified in at least 2 of 3 independent whole-genome scans*

Chromosome location	Data set		
	UK	ECRAF	NARAC
1q	*D1S238* (203 cM) ($P = 0.05$)	*D1S238* (203 cM)† ($P = 0.022$)	*D1S235* (259 cM) ($P = 0.005$)
3q		*D3S1267* (139 cM)‡ ($P = 0.039$)	*D3S4523* (138 cM)‡ ($P = 0.029$)
6q	*D6S434* (109 cM) ($P = 0.0007$)		*D6S1021* (112.2 cM) ($P = 0.008$)
8p		*D8S1825* (15.4 cM)† ($P = 0.040$)	*D8S277* (8.3 cM) ($P = 0.009$)
12q		*D12S95* (96.1 cM) ($P = 0.043$)	*D12S1052* (83.2 cM) ($P = 0.023$)
14q	*D14S275* (28 cM) ($P = 0.03$)		*D14S1280* (25.9 cM) ($P = 0.017$)
14q	*D14S276* (56.4 cM) (P=0.05)	*D14S285* (59.4 cM)† ($P = 0.049$)	*D14S578* (55.8 cM) ($P = 0.037$)
16p		*D16S420* (44.5 cM) ($P = 0.039$)	*D16S403* (43.9 cM) ($P = 0.004$)
18q		*D18S474* (71.3 cM) ($P = 0.012$)	*D18S858* (80.4 cM) ($P = 0.043$)

* All results were obtained by multipoint linkage analysis unless indicated otherwise. Distance from HLA-DRB1 is shown in parentheses. Marker positions obtained from Marshfield map (sex-averaged positions) (www.marshfieldclinic.org/research/genetics). ECRAF = European Consortium on Rheumatoid Arthritis Families; NARAC = North American Rheumatoid Arthritis Consortium.
† By single-point analysis.
‡ Insulin-dependent diabetes mellitus 9 locus.

important to independently validate all findings in further data sets. To date, at least 2 studies have demonstrated nominal linkage to overlapping regions on chromosomes 3q, 8p, 12q, 14q, 16p, 16q, 18q, and Xp (8–10) (Tables 1 and 2). These regions are important candidates for containing true RA susceptibility loci, but because the individual RA whole-genome scans undertaken to date had relatively low power to identify linkage to genes of small effect, some regions of true linkage are likely to have been missed.

We now describe a systematic whole-genome scan and affected–sibling pair linkage analysis, which was performed on a large, well-characterized cohort of RA families in the UK in order to identify novel susceptibility loci and confirm linkage to previously identified regions.

PATIENTS AND METHODS

RA-affected sibling pairs. The study group comprised 252 Caucasian sibling pairs affected by RA. The affected sibling pairs, from 182 families, were identified from The Arthritis and Rheumatism Council's UK National Repository of family material (http://www.arc.man.ac.uk) (11). All family members were examined according to a standard protocol and given a detailed, structured interview regarding joint symptoms. A trained metrologist performed a joint examination to detect any swelling, tenderness, or deformity. Hand radiographs were reviewed, and rheumatoid factor (RF) status was ascertained using a particle agglutination test. Subjects were classified as having RA if they satisfied the 1987 American College of Rheumatology (formerly, the American Rheumatism Association) criteria (12) modified for genetic studies (13).

DNA from the affected individuals and their first-degree relatives was prepared from peripheral blood. Parental DNA was available in 37% of the families. In 94% of the remaining families, DNA from unaffected siblings was available, facilitating the assignment of parental genotypes. In 26 families, 3 or more siblings were affected by disease. Of the 650 individuals included in the study, 403 had RA (393 siblings, 10 mothers, and 3 fathers; 3 of the affected parents were also members of an affected sibling pair). Seventy-six percent of affected individuals were female, 85% were seropositive for RF, and 80% had radiographically apparent hand erosions. The mean (±SD) age at disease onset was 39.4 ± 13.2 years, and the mean disease duration at recruitment was 15.3 ± 11.3 years. In 60% of cases, both affected siblings were female, both were male in 7%, and the other cases comprised mixed-sex pairs.

HLA–DR status. HLA–DRB1 genotypes were determined using a commercially available semiautomated polymerase chain reaction (PCR)–sequence-specific oligonucleotide probe typing technique (INNO-LiPA; Abbott Laboratories, Maidenhead, UK).

Microsatellite markers. Three hundred sixty-five highly informative microsatellites from the ABI PRISM Link-

Table 2. Loci detected by single-point linkage analysis in 252 affected sibling pair families with rheumatoid arthritis in the UK*

Chromosome location	Data set, marker (position)			
	UK	ECRAF	NARAC	Other studies/diseases†
3	*D3S2338* (42 cM)		*D3S3038* (44.8 cM)	
4	*D4S1592* (69.5 cM)			
7	*D7S484* (53.3 cM)			*D7S484*/ MS + CD/UC + asthma
10	*D10S185* (116.3 cM)		*D10S2470* (112.6 cM)	
	D10S192 (124 cM)			*D10S217* (167 cM)/IDDM
	D10S217 (158 cM)			
16	*D16S3103* (32 cM)	*D16S420* (42.5 cM)	*D16S403* (43.9 cM)	*Pgia3*‡/ IBD1§
		D16S401 (47 cM)		
21	*D21S1256* (9.7 cM)			
X	*DXS1106* (66.6 cM)			

* Results are from whole-genome scans. Marker positions were obtained from Marshfield map (sex-averaged positions) at www.marshfieldclinic.org/research/genetics. MS = multiple sclerosis; CD/UC = Crohn's disease/ulcerative colitis; IBD = inflammatory bowel disease (see Table 1 for other definitions).
† Studies of related autoimmune diseases with linkage to the region. Linkage results are summarized in ref. 29.
‡ Refs. 20 and 21.
§ Ref. 22.

age Mapping Set version 2 marker set (Applied Biosystems, Warrington, UK) were genotyped. The microsatellites spanned the whole genome, with a mean distance between markers of ~10 cM and mean heterozygosity of 0.78.

Microsatellite genotyping. A semiautomated analysis of microsatellite genotypes was carried out in 2 centers; chromosomes 1–7, 16, 21, and 22 were typed at the Wellcome Trust Centre for Human Genetics (WTCHG), and chromosomes 8–15, 17–20, and X were typed at the Arthritis Research Campaign Epidemiology Unit (ARC-EU). The microsatellites were amplified by PCR using fluorescently labeled primers, and reactions were performed in 10-μl volumes (each containing 50 ng of DNA, 5 pmoles of each PCR primer, 4 nmoles of each of the 4 deoxynucleotide triphosphates, and 0.2 units of *Taq* polymerase [Bioline, London, UK] in 1–3 m*M* $MgCl_2$ buffer, overlaid with liquid paraffin). The reactions were performed in 96-well microtiter plates with 30–35 cycles of denaturation (1 minute at 95°C), primer annealing (30 seconds at 55–60°C), and extension (30 seconds at 72°C). Annealing temperatures and $MgCl_2$ buffer concentration were optimized for each primer.

Amplification of each microsatellite was performed separately, and the PCR products were then combined into pools of 10–20 markers before loading. Products were separated by electrophoresis on either 6% polyacrylamide gels using ABI 373 semiautomated DNA sequencers (Applied Biosystems) over 3 hours (WTCHG) or 4% polyacrylamide gels using ABI 377 DNA sequencers over 2 hours (ARC-EU). Products were sized using the program GeneScan 672, version 2.1 (Applied Biosystems), and genotypes were semiautomatically assigned using the program Genotyper, version 1.1.1 (Applied Biosystems). A PCR product from a DNA reference sample (Centre d'Etudes du Polymorphisme Humain [CEPH] 1347-02) was included on every gel to monitor possible gel-to-gel variation.

Statistical analysis. *Error checking.* To minimize data errors, extensive error-checking procedures were used. Allele assignment by Genotyper was checked manually for all genotypes, and the CEPH control sample was used to ensure the consistency of allele assignment. Size data were converted into discrete allele numbers, and samples not following Mendelian patterns of inheritance were identified in PedCheck (14) and removed from the data set. The family data were then assessed using RELATIVE software (15). RELATIVE identifies probable monozygotic twins, half-siblings, or unrelated individuals by testing whether the proportion of alleles sharing identity by descent at unlinked loci (on the basis of ≥50 markers) is consistent with the expected proportion for each relative pair.

Inter- and intraobserver variation. To assess interobserver variation, at both centers a panel of samples was genotyped for 6 microsatellites on chromosome 7. An additional 10 markers were regenotyped in a blinded manner by those researchers undertaking the genotype analysis. Intraobserver variation was then determined.

Nonparametric analysis. Allele frequencies were calculated from all of the scored genotypes using the software

package DOWNFREQ (Terwilliger J: unpublished) (WTCHG) or SPLINK, version 1.05 (David Clayton, MRC Biostatistics Unit, Cambridge, UK) (ARC-EU). Sibling pair analysis methods were then used to test for excess allele sharing between affected siblings. Nonparametric single-point and multipoint analysis was implemented in MAPMAKER/SIBS, version 2 (16). Multiple sibships were given a conservative weighting of 2/n to account for their lack of independence (17). Analysis of markers on the X chromosome was carried out using MAPMAKER/SIBS modified for X-linked loci (18). The program calculates sharing between affected sister–sister, brother–brother, and sister–brother sib pairs.

Both multipoint and single-point analyses are presented. We used the suggested thresholds of logarithm of odds (LOD) scores ≥ 3.6 ($P \leq 2 \times 10^{-5}$) for significant linkage and ≥ 2.2 ($P \leq 7.4 \times 10^{-4}$) for suggestive linkage (18). Nominal evidence of linkage was based on an LOD score ≥ 0.8 ($P < 0.05$) by either multipoint or single-point analysis.

RESULTS

Linkage analysis. The results of the nonparametric single-point and multipoint analyses implemented in MAPMAKER/SIBS are shown in Figure 1. Linkage to the HLA region on chromosome 6 was confirmed by multipoint analysis (maximum LOD score 4.7 at marker *D6S276*, $P = 0.000003$; 1 cM from HLA–DRB1). The allele-sharing ratio for DRB1 was 16:49:35 for 0-, 1-, and 2-allele sharers. This significant increase in the inheritance of 2 alleles identical by descent provides evidence for linkage to DRB1 (single-point LOD score 1.9). Suggestive evidence of linkage by multipoint analysis was identified on chromosome 6q (*D6S434*, $P = 0.0007$) at 109 cM, and nominal evidence of linkage by multipoint analysis was found for 3 additional linkage intervals: 1q (*D1S2842* and *D1S2836*, $P = 0.05$) at 277–290 cM, 14q (*D14S283*, $P = 0.05$, and *D14S275*, $P = 0.03$) at 28 cM, and 14q (*D14S276*, $P = 0.03$) at 56.4 cM. Single-point analysis also identified suggestive linkage for 1 marker on chromosome 6q (*D6S434*, $P = 0.0006$) at 109 cM, and nominal linkage ($P < 0.05$) was identified for 10 additional linkage intervals on chromosomes 3p (*D3S2338* at 42 cM), 4q (*D4S1592* at 69.5 cM), 7p (*D7S484* at 53.5 cM), 10q (*D10S192* at 124 cM), 10q (*D10S217* at 158 cM), 14q (*D14S275* at 28 cM), 14q (*D14S276* at 56.4 cM), 16p (*D16S3103* at 32 cM), 21q (*D21S1256* at 9.7 cM) and Xq (*DXS1106* at 66.6 cM).

Comparisons with data from the European and North American whole-genome screens are shown in Tables 1 and 2.

Microsatellite markers. The microsatellite markers spanned the genome, with a mean distance between markers of 10 cM and a median distance of 9.3 cM. Eighty-eight percent of markers were within 14 cM of each other. Only 11 markers had a gap of >20 cM, using the Marshfield map, as a result of marker failure. The mean (±SD) heterozygosity was 0.78 ± 0.07, and the mean polymorphism information content was 0.75 ± 0.08.

Error checking. Fifteen individuals whose DNA samples did not follow Mendelian inheritance patterns were identified by PedCheck (14) and removed from the analysis. Three previously unrecognized half-siblings and 2 other individuals (whose samples did not fit Mendelian inheritance patterns) were identified by RELATIVE (15) and were also removed from the analysis.

Inter- and intraobserver variation. For the 6 microsatellite markers on chromosome 7 that were genotyped at both centers, concordance was 98.9%. Intraobserver variation, measured by regenotyping 10 markers in a blinded manner, ranged from 0% to 3%.

DISCUSSION

Consistent with the results of 2 previously reported RA whole-genome screens (9,10), this study has confirmed genetic linkage to the HLA region on chromosome 6. In addition, suggestive evidence of linkage ($P < 7.4 \times 10^{-4}$) to a non-HLA region on 6q was identified by both single- and multipoint analyses. Nominal evidence of linkage ($P < 0.05$) to 10 other regions (chromosomes 3p, 4q, 7p, 2 regions on 10q, 2 regions on 14q, 16p, 21q, and Xq) was detected by single-point analysis, and to 3 linkage intervals (chromosomes 1q, and 2 regions on 14q) by multipoint analysis.

Regions identified as nominal-linkage intervals by 2 or more whole-genome scans are of greater interest than are sites identified by only 1 study, especially when corroborative evidence from other autoimmune diseases or animal studies is available. Six of the non-HLA regions identified in this study were also reported by the ECRAF (9) and/or NARAC (10) studies (chromosomes 3q, 6q, 10q, 14p, 14q, and 16p), and several of them overlap with regions linked to other autoimmune diseases (Tables 1 and 2). The non-HLA region on 6q with suggestive evidence of linkage was also identified in the NARAC study (10). On chromosome 10, the linkage interval including *D10S185* (116 cM) and *D10S192* (124 cM) was identified in the NARAC study (10) and has also been identified in an ankylosing spondylitis whole-genome scan (20,21). In addition, the interval around *D10S217* (167 cM) has been identified in a recent second-generation whole-genome scan in insulin-dependent diabetes mellitus (IDDM) (22). The region

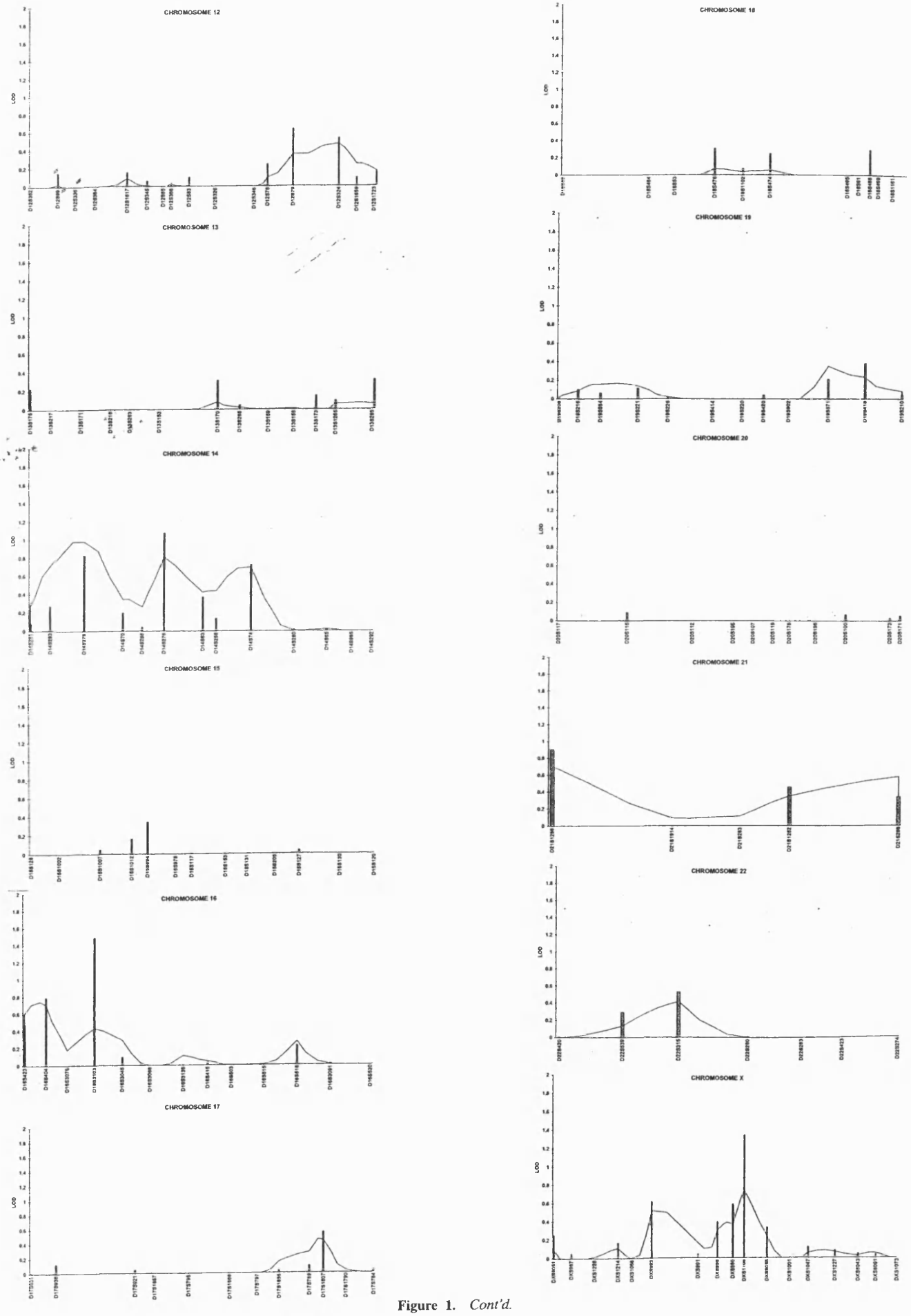


Figure 1. *Cont'd.*

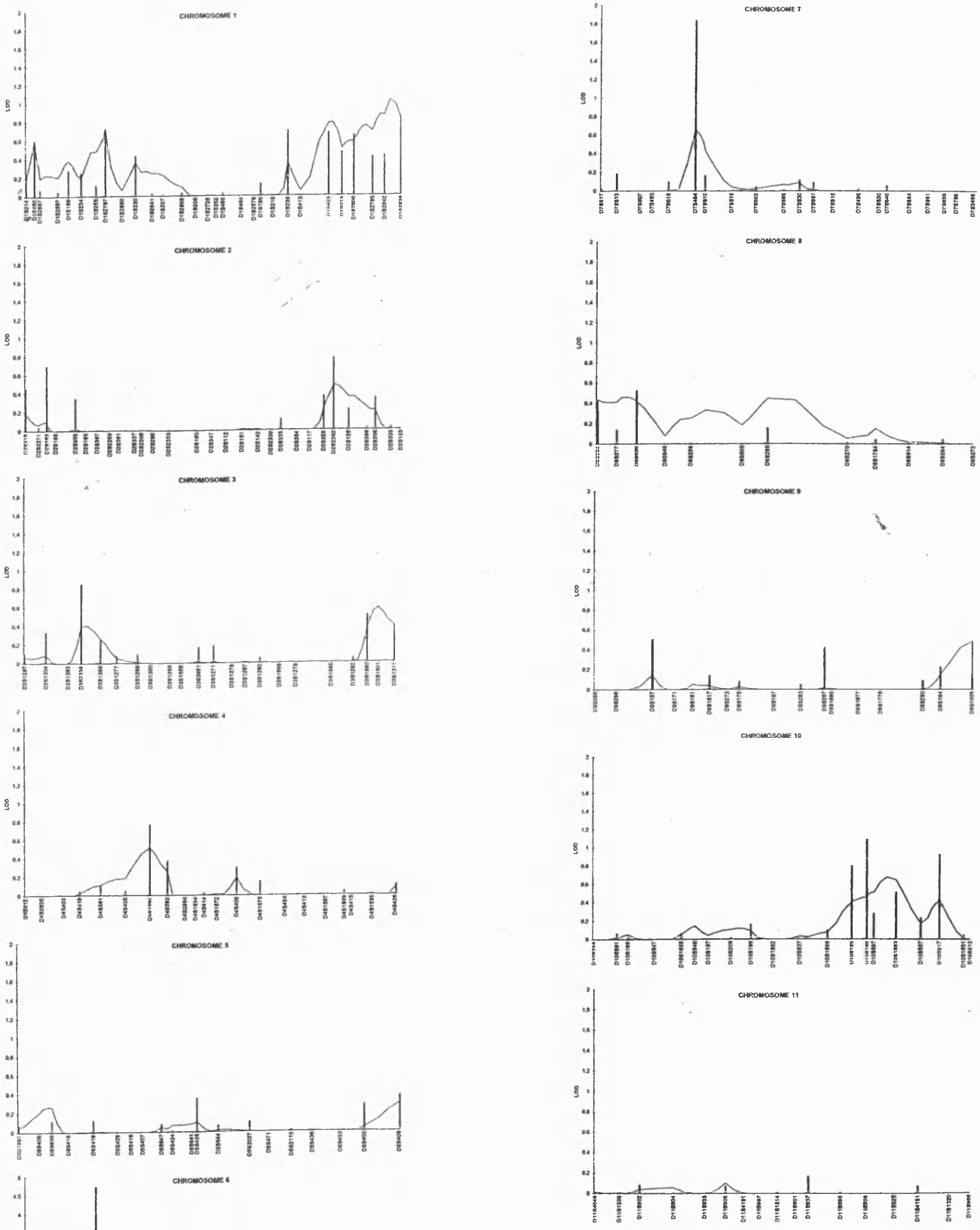


Figure 1. Results of single- and multipoint linkage analysis for all chromosomes, obtained using MAPMAKER/SIBS. Bars represent the results of single-point linkage analysis. The continuous lines represent the results of the multipoint linkage analysis. The logarithm of odds (LOD) score is on the y axis, and for all graphs except that for chromosome 6, the scale is the same (maximum LOD score 2). Marker names are on the x axis.

on 14q including *D14S276* (56 cM) was identified by both the ECRAF (9) and NARAC (10) studies, as was the interval on 16p (*D16S3103*). Recently, a whole-genome scan conducted using mice with proteoglycan-induced arthritis (23) also identified a linkage interval (named *Pgia3*) homologous to this same region on 16p.

In accordance with findings in the NARAC study (10), we are unable to provide supporting evidence for the IDDM 6 locus on chromosome 18, a site that showed evidence for linkage in the ECRAF study (9). Previous investigations using some of the same RA affected sibling pair families as in this study reported linkage to 3 candidate genes and are described below. The current study did not attempt to replicate these findings.

Using a number of single-nucleotide polymorphisms and microsatellite markers spanning the interleukin-1 gene cluster, Cox et al detected some evidence of linkage using the combined transmission disequilibrium test (TDT) and sib-TDT in a subset of families sharing only 1 or 0 DRB1 alleles identical by descent. However, they found no evidence of linkage using nonparametric allele-sharing methods (24). John et al reported linkage to a single marker mapping to an intron within the estrogen synthase gene (CYP19) (25). In the current study, the closest marker mapped 5 cM from CYP19, which may explain the failure to replicate the previous finding.

Fife et al investigated corticotropin-releasing hormone (CRH) as a candidate RA susceptibility locus (26). Linkage to a 10 cM region adjacent to CRH was reported. No statistically significant evidence of linkage was detected in the current study, although there was some overlap of the families used in the 2 investigations. It is interesting to note, however, that there is an apparent small peak on multipoint linkage analysis to the same region in the current study (Figure 1). With the marker density used in this study, linkage to these (and other postulated disease) loci could have been missed and cannot, therefore, be excluded as potential disease genes.

Power calculations suggest that the current study should have 80% power to detect loci contributing $\lambda = 1.6$, using the threshold for detection as a LOD ≥ 1.0 (27). This, in fact, may be an overestimate, because power calculations typically assume that all families are fully informative (i.e., DNA is available from all parents, and marker heterozygosity is 100%), and that all affected sibling pairs are independent (which is not the case in families with >2 affected siblings). Recruitment of complete affected sibling pair families in a late-onset disease such as RA is difficult. Thus, parental DNA was available in just over one-third of our families. However, one of the strengths of this study is that DNA from unaffected siblings was available in 94% of the remaining families, and this was used to infer parental haplotypes in families in which they were not directly available.

Because of the high chance of false-positive results in a genome-wide linkage study, we applied the stringent thresholds for significance ($P < 2.2 \times 10^{-5}$) recommended by Lander and Kruglyak (18). However, if stringent thresholds for significance are interpreted too strictly in a disease such as RA, with a relatively low λ_s, it would be easy to overlook potentially relevant evidence emerging from whole-genome scans.

In order to make best use of the information arising from this and similar-sized whole-genome scans, a number of strategies are available. One proposal is to perform the initial screen in a relatively small data set and then to test those loci demonstrating nominal evidence of linkage ($P < 0.05$) in a larger data set (28). In this respect, our study can be seen as a replication data set in which to test positive linkages detected in the US and European whole-genome scans. As outlined above, a number of positive linkages from the NARAC and ECRAF studies have also shown evidence of linkage in our data set, suggesting that true RA susceptibility genes may map to these regions. A complementary but alternative approach would be to undertake a meta-analysis of the genotype data in previously published whole-genome screens. By combining these data, it should be possible to make more robust interpretations of the evidence, thereby focusing the efforts of the replication studies.

In conclusion, our results of the UK whole-genome linkage analysis of RA susceptibility loci confirm significant evidence of linkage in the HLA region by single-point and multipoint analysis. Suggestive linkage was identified by single- and multipoint analysis at a non-HLA region on 6q. Nominal evidence of linkage ($P < 0.05$) was found by single-point analysis for 10 other regions on chromosomes 3p, 4p, 7p, 2 regions of 10q, 2 regions of 14q, 16p, 21p, and Xq, and by multipoint analysis for 3 non-MHC regions on chromosomes 1q and 2 regions on 14q.

REFERENCES

1. Silman AJ, MacGregor AJ, Thomson W, Holligan S, Carthy D, Farhan A, et al. Twin concordance rates for rheumatoid arthritis: results from a nationwide study. Br J Rheumatol 1993;32:903–7.
2. Aho K, Koskenvuo M, Tuominen J, Kaprio J. Occurrence of

rheumatoid arthritis in a nationwide series of twins. J Rheumatol 1986;13:899–902.
3. Deighton CM, Walker DJ. The familial nature of rheumatoid arthritis. Ann Rheum Dis 1991;50:62–5.
4. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum 2000;43:30–7.
5. Ollier W, Thomson W. Population genetics of rheumatoid arthritis. Rheum Dis Clin North Am 1992;18:741–59.
6. Deighton CM, Walker DJ, Griffiths ID, Roberts DF. The contribution of HLA to rheumatoid arthritis. Clin Genet 1989;36: 178–82.
7. Wordsworth BP. HLA class II antigens in susceptibility to rheumatoid arthritis. Br J Rheumatol 1991;30:151–2.
8. Shiozawa S, Hayashi S, Tsukamoto Y, Goko H, Kawasaki H, Wada T, et al. Identification of the gene loci that predispose to rheumatoid arthritis. Int Immunol 1998;10:1891–5.
9. Cornelis F, Faure S, Martinez M, Prud'homme JF, Fritz P, Dib C, et al. New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. Proc Natl Acad Sci U S A 1998;95: 10746–50.
10. Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Monteiro J, et al. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. Am J Hum Genet 2001;68:927–36.
11. Worthington J, Ollier WER, Leach MK, Smith I, Hay EM, Thomson W, et al. The Arthritis and Rheumatism Council's National Repository of Family Material: pedigrees from the first 100 rheumatoid arthritis families containing affected sibling pairs. Br J Rheumatol 1994;33:970–6.
12. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
13. MacGregor AJ, Bamber S, Silman AJ. A comparison of the performance of different methods of disease classification for rheumatoid arthritis: results from an analysis from a nationwide twin study. J Rheumatol 1994;21:1420–6.
14. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 1998;63:259–66.
15. Goring HH, Ott J. Relationship estimation in affected sib pair analysis of late-onset diseases. Eur J Hum Genet 1997;5:69–77.
16. Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 1995;57: 439–54.
17. Meunier F, Philippi A, Martinez M, Demenais F. Affected sib-pair tests for linkage: type I errors with dependent sib-pairs. Genet Epidemiol 1997;14:1107–11.
18. Cordell HJ, Kawaguchi Y, Todd JA, Farrall M. An extension of the Maximum Lod Score method to X-linked loci. Ann Hum Genet 1995;59:435–49.
19. Lander ES, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995;11:241–7.
20. Brown MA, Pile KD, Kennedy LG, Campbell D, Andrew L, March R, et al. A genome-wide screen for susceptibility loci in ankylosing spondylitis. Arthritis Rheum 1998;41:588–95.
21. Laval SH, Timms A, Edwards S, Bradbury L, Brophy S, Milicic A, et al. Whole-genome screening in ankylosing spondylitis: evidence of non-MHC genetic-susceptibility loci. Am J Hum Genet 2001; 68:918–26.
22. Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, et al. A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. Nat Genet 1998;19:292–6.
23. Otto JM, Cs-Szabo G, Gallagher J, Velins S, Mikecz K, Buzas EI, et al. Identification of multiple loci linked to inflammation and autoantibody production by a genome scan of a murine model of rheumatoid arthritis. Arthritis Rheum 1999;42:2524–31.
24. Cox A, Camp NJ, Cannings C, di Giovine FS, Dale M, Worthington J, et al. Combined sib-TDT and TDT provide evidence for linkage of the interleukin-1 gene cluster to erosive rheumatoid arthritis. Hum Mol Genet 1999;8:1707–13.
25. John S, Myerscough A, Eyre S, Roby P, Hajeer A, Silman AJ, et al. Linkage of a marker in intron D of the estrogen synthase locus to rheumatoid arthritis. Arthritis Rheum 1999;42:1617–20.
26. Fife MS, Fisher SA, John S, Worthington J, Shah CJ, Ollier WE, et al. Multipoint linkage analysis of a candidate gene locus in rheumatoid arthritis demonstrates significant evidence of linkage and association with the corticotropin-releasing hormone genomic region. Arthritis Rheum 2000;43:1673–8.
27. Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 1990;46:229–41.
28. Holmans P, Craddock N. Efficient strategies for genome scanning using maximum-likelihood affected-sib-pair analysis. Am J Hum Genet 1997;60:657–66.
29. Becker KG, Simon RM, Bailey-Wilson JE, Freidlin B, Biddison WE, McFarland HF, et al. Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. Proc Natl Acad Sci U S A 1998;95:9979–84.

Rheumatology 2003;42:149–153
doi:10.1093/rheumatology/keg054, available online at www.rheumatology.oupjournals.org

Rheumatoid arthritis susceptibility and interleukin 10: a study of two ethnically diverse populations

K. MacKay, A. Milicic, D. Lee, M. Tikly[1], S. Laval, J. Shatford and P. Wordsworth

Introduction. **IL-10 is an immunoregulatory cytokine which may modulate disease expression in rheumatoid arthritis (RA). The IL-10 gene is highly polymorphic with a number of single nucleotide polymorphisms in the promoter region and two microsatellite loci, IL10.R and IL10.G, 4 kb and 1.1 kb 5′ of the transcription initiation site. It has been reported that allele 2 of the IL10.R microsatellite (IL10.R2) is associated with increased IL-10 secretion and IL10.R3 with reduced secretion. Subsequently, over-representation of IL10.R2 and under-representation of IL10.R3 in three independent RA groups has been reported. The aim of the current study is to determine whether there is an association between the IL10.R2 allele and RA in two ethnically distinct populations.**

Methods. **IL10.R genotypes were determined by semi-automated DNA sequencing technology in 186 UK Caucasians and 138 South Africans of Zulu or Sotho origin, fulfilling the 1987 American College of Rheumatology (ACR) criteria for RA. The Caucasian patients had relatively severe disease and comprised 75 patients with RA vasculitis, 22 with Felty's syndrome and 89 who had undergone a joint replacement (hip or knee) within 15 years of the onset of disease. Allele frequencies were compared with 296 Caucasians and/or 73 South Africans.**

Results. **The frequency of the IL10.R2 allele was significantly greater in the South Africans (RA and controls) than in the Caucasians (0.78 *vs* 0.66, $P=1\times10^{-6}$), while the frequency of IL10.R3 was less common (0.16 *vs* 0.3, $P=1\times10^{-8}$). No differences were observed in either IL10.R2 or IL10.R3 frequencies between patients and controls in either population.**

Conclusions. **We were unable to confirm any association between IL10.R alleles and RA in this study. However, significant differences were demonstrated in the frequency of IL10.R2 and IL10.R3 between the two ethnic groups. The relatively high frequency of IL10.R2 in the South African population (0.78) would have reduced the power to detect an association with RA.**

KEY WORDS: Rheumatoid arthritis, Interleukin 10, Disease susceptibility, Ethnic diversity.

Rheumatoid arthritis (RA) is a common systemic inflammatory arthropathy characterized by chronic synovitis and progressive joint destruction. Pro-inflammatory cytokines, such as tumour necrosis factor (TNF) and interleukin 1 (IL-1), play a significant pathogenic role and the disease can be ameliorated by treatments specifically targeting TNF [1] or IL-1 [2]. Anti-inflammatory cytokines can also be found in the affected joints and it has been postulated that chronic synovitis may reflect an imbalance in pro- and anti-inflammatory cytokine

Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford, UK and [1]Chris Hani Baragwanath Hospital and University of the Witwatersrand, PO Bertsham 2013, South Africa.

Submitted 25 July 2001; revised version accepted 20 June 2002.

Correspondence to: P. Wordsworth, Nuffield Orthopaedic Centre, Windmill Road, Headington, Oxford OX3 7LD, UK. E-mail: paul.wordsworth@clinical-medicine.oxford.ac.uk

production in RA [3]. Consequently, the potential beneficial effects of anti-inflammatory cytokines such as interleukin 10 (IL-10) [4] and interleukin 4 (IL-4) [5] in RA are of great interest.

The immunomodulatory cytokine IL-10 is produced by a variety of cell types, including monocytes [6] and B lymphocytes [7]. It is a potent up-regulator of B-cell production and differentiation [8], but has anti-inflammatory capabilities that can directly down-regulate TNFα, IL-1, IL-8 and interferon-γ production [6, 9]. Variation in IL-10 secretion is largely genetically determined [10] and differences in secretion have been associated with various chronic inflammatory and infectious diseases. High levels of secretion are associated with a poor or fatal outcome in meningitis [10] and low levels are associated with susceptibility to severe malarial anaemia [11].

The IL-10 gene maps to the junction of 1q31–q32 [12] and exhibits substantial polymorphism in the promoter region which appears to correlate with variation in transcription [13, 14]. Two microsatellite polymorphisms, IL10.G and IL10.R, situated 1.1 and 4 kb 5′ of the transcription initiation site, respectively, are of particular interest [15, 16] since variation in IL-10 secretion has been associated with particular haplotypes defined by these microsatellite markers. Haplotypes containing the IL10.R2 allele are associated with higher levels of secretion than those including IL10.R3 [17]. Eight single nucleotide polymorphisms (SNPs) have also been identified in the promoter region of this gene [18–20], three of which have been studied in some detail: −1082 (G to A), −819 (C to T) and −592 (C to A) [13, 18, 20]. Increased IL-10 secretion has been described with the common GCC haplotype and reduced IL-10 secretion with the least common ATA haplotype, but direct comparisons between these studies are difficult as they have employed different experimental protocols [13, 20].

Various IL10.R, IL10.G and SNP genotypes have been reported to show association with a variety of chronic inflammatory diseases, including the ATA haplotype with extended oligo-articular juvenile idiopathic arthritis [13], IL10.G with systemic lupus erythematosus [21] and recently the IL10.R2 allele with RA [22].

This study was undertaken to investigate further the association of IL10.R2 with RA in UK Caucasians and to determine if it extended to another racially distinct population.

Materials and methods

Patients and controls

Two groups of racially distinct patients with RA and ethnically matched controls were recruited from the United Kingdom and South Africa. The patients consisted of 186 Caucasian and 138 South African (SA) patients of Sotho or Zulu ethnicity, fulfilling the American College of Rheumatology (ACR) 1987 criteria for RA. The Sotho and Zulu groups have a similar genetic background [23]. All the UK Caucasian patients had relatively severe disease and comprised 75 patients with RA vasculitis, 22 with Felty's syndrome and 89 who had undergone a large joint replacement (hip or knee) within 15 years of the onset of disease. The mean age (±S.D.) of the Caucasian RA patients was 64 (±11.6) yr and 70 per cent were female. Mean disease duration was 18.7 (±10.5) yr and mean age of disease onset was 44.9 (±14.5) yr. The SA patients were not chosen specifically for disease severity but were all recruited from a tertiary hospital out-patient clinic and required disease-modifying therapy. The mean age (±S.D.) of the SA patients was 49.5 (±10.9) yr, average disease duration was 7.6 (±7.3) yr and 76 per cent were female. The UK Caucasian control group consisted of 210 healthy blood donors and 86 unaffected spouses of patients attending a skeletal dysplasia clinic. The South African controls included 73 healthy, ethnically matched hospital workers or hospital out-patients attending with minor trauma.

IL10.R genotyping

Genomic DNA was extracted from peripheral venous blood samples using standard techniques. The IL10.R microsatellite was amplified by polymerase chain reaction (PCR) [primers were 5′ CCC TCC AAA ATC TAT TTG CAT A (upstream) and 5′ CTC CGC CCA GTA AGT TTC ATC (downstream), the latter being tagged with a fluorescent dye (HEX)]. Reactions were performed in 96-well plates (Costar) in 10 ml reactions consisting of 50 ng DNA, 400 nM each primer, 50 mM each dNTP, 2.0 mM $MgCl_2$ and 0.2 units DNA polymerase (Bioline, UK) in the manufacturer's NH_4 buffer. The cycling conditions were 94°C for 1 min, annealing 60°C for 1 min, extension 72°C for 1 min, for 32 cycles. PCR products were diluted with water and separated by electrophoresis using an ABI 373 semi-automated sequencer (Applied Biosystems, Warrington, UK) and 6% denaturing polyacrylamide gels over 3 h. Products were sized using the program GENESCAN™ Version 2.1 (Applied Biosystems, Warrington, UK) and genotypes semi-automatically assigned using the program GENOTYPER™ Version 1.1 (Applied Biosystems, Warrington, UK). All genotypes were then verified manually. The program GAS (Version 2) (A. Young, unpublished) was used to convert the size data into discrete allele numbers.

HLA-DR typing

Sequence-specific PCR, using 35 primers, was used to differentiate between the different HLA-DR alleles and undertake DR4 and DR1 subtyping [24].

Statistical analysis

Allele and genotype frequencies were calculated by direct counting. Since a common source of error in genotype assignment is the over-calling of homozygotes, Hardy–Weinberg equilibrium was used to predict the likely frequencies of IL10.R2 and IL10.R3 homozygotes and these figures were compared with the observed frequencies. IL10.R allele frequency distribution was compared between the Caucasians and South Africans and then between patients and controls. Subgroup analysis included disease severity (RA vasculitis, Felty's syndrome or an early large joint replacement), sex, and shared epitope status (homozygosity or heterozygosity). The significance of differences between groups was calculated from contingency tables by χ^2 analysis. Odds ratios with confidence intervals were calculated.

Results

The frequency of the IL10.R2 allele was significantly higher ($P = 1 \times 10^{-6}$) in the South African population

overall (0.78) than the UK Caucasians (0.66), while the frequency of IL10.R3 was correspondingly reduced (0.16 *vs* 0.30, $P=1\times10^{-8}$). However, no differences were observed in IL10.R allele frequencies between patients and controls in either racial group (Table 1). Eighty-four per cent of the UK Caucasian patients and 58 per cent of the SA patients were positive for the shared epitope.

A variety of subgroups were defined from the UK Caucasian RA cohort to analyse any possible associations with IL10.R alleles. All allele frequencies were very similar whether the groups were divided by sex, age of onset, shared epitope status, extra-articular disease (RA vasculitis and Felty's syndrome) or early large joint replacement. The observed frequencies of IL10.R2 and IL10.R3 homozygous genotypes compared well with the predicted frequency of homozygotes by Hardy–Weinberg equilibrium suggesting alleles were being appropriately assigned.

TABLE 1. *Frequencies of IL10.R alleles in the South African and Caucasian populations studied*[a]

Comparison groups	IL10.R alleles	Alleles		Alleles		*P*	Odds ratio (C.I. 95%)
		Caucasian		South African			
		No.	%	No.	%		
Caucasians *vs* South Africans	IL10.R1	4	0.4	11	2.5	ND	
	IL10.R2	640	66	350	79	1×10^{-6}	
	IL10.R3	287	30	69	16	1.4×10^{-8}	
	IL10.R4	32	3.3	11	2.6	0.4	
	IL10.R5	1	0.1	1	0.3	0.6	
		Patients		Controls			
		No.	%	No.	%		
Caucasian patients *vs* controls	IL10.R1	3	0.8	1		0.2	
	IL10.R2	253	68	387	65	0.4	1.3 (0.8–1.5)
	IL10.R3	103	28	184	31	0.2	0.85 (0.6–1.4)
	IL10.R4	13	2.7	19	3.2	0.8	
	IL10.R5	0	0	1	0.2	ND	
South African patients *vs* controls	IL10.R1	9	3.2	2	1.4	ND	
	IL10.R2	223	78	127	77	0.4	0.8 (0.5–1.4)
	IL10.R3	47	17	22	15	0.5	1.25 (0.7–2.2)
	IL10.R4	7	2.5	4	2.7	ND	
	IL10.R5	0	0	1	0.7	ND	

[a]Allele frequencies were compared using χ^2 analysis.
ND, not done because numbers were too small for reliability.

TABLE 2. IL10.R2 and IL10.R3 allele frequencies in RA patients (from different ethnic backgrounds) *vs* controls using data from this study and the study undertaken by Eskdale *et al.*

Study	Patient group	IL10.R2 % RA	IL10.R2 % control	IL10.R3 % RA	IL10.R3 % control
This study	Caucasians with severe RA RA ($n=186$)				
	Control ($n=296$)	68	65	28	31
This study	Black South Africans (DSND) RA ($n=138$)				
	Control ($n=73$)	78	77	17	15
Lancet 1998 (Eskdale *et al.*)	Oxford Caucasians (DSND) RA ($n=148$)				
	Control ($n=87$)	70	61	25	36
Lancet 1998 (Eskdale *et al.*)	Glasgow Caucasians (DSND) RA ($n=103$)				
	Control ($n=94$)	69	56	29	40
Lancet 1998 (Eskdale *et al.*)	African Americans (DSND) RA ($n=61$)				
	Control ($n=38$)	87	72	11	24

DSND, disease severity not documented.

Discussion

No association between IL-10 alleles and RA was apparent in the current study, in contrast to previous reports of an increase in IL10.R2 and a reduction in IL10.R3 alleles [22]. There are several possible explanations for this. First, the previous reported association with the IL10.R2 allele may be spurious. Second, it may be relevant that the IL10.R2 allele frequency in UK Caucasian controls in our study is significantly higher (0.65 *vs* 0.59, $P=0.03$) than that reported by Eskdale *et al.* and the IL10.R3 frequency significantly lower (0.31 *vs* 0.38, $P=0.02$). No differences in allele frequencies of IL10.R2 or IL10.R3 were apparent when the UK Caucasian patient groups from the two studies were compared (Table 2). This may suggest that IL-10 contributes a weak genetic effect but the relatively high frequency of IL10.R2 in the general population makes the effect more difficult to detect reliably. Since the frequency of IL10.R2 is even higher in South African Sotho and Zulu populations, the power to detect association with RA would be further reduced. Previously reported estimates of relative risk for IL10.R in RA range between 1.5 and 2.5 [22]. We estimate that our study in UK Caucasians had 80 per cent power to exclude an association between IL-10 and RA with an odds ratio of $\geqslant 1.8$. However, in the South African study the higher frequency of the IL10.R2 allele (SA control frequency 77 *vs* 65% in Caucasians) means that the power to detect an effect with an odds ratio of $\geqslant 1.8$ was only 36 per cent. In contrast the study had 80 per cent power to exclude a putative genetic effect with an odds ratio of $\geqslant 3$. Third, genetic heterogeneity may be operating. The UK patients in this study were specifically selected for having more severe forms of RA and were only included if they had undergone an early large joint replacement or fulfilled the criteria for rheumatoid vasculitis or Felty's syndrome. This is in contrast to the previous report where the main recruitment criteria was RA fulfilling the 1987 ACR criteria [22]. It is therefore likely that the disease severity of the Caucasian patients included in the two studies was different and it is conceivable that the IL10.R2 allele is not as strongly associated with severe forms of RA as it is with milder variants.

Inter-ethnic differences in IL-10 allele frequencies were not unexpected as similar differences have been described previously for the TNF locus [25–27] although in the case of TNF, linkage disequilibrium within the MHC may contribute to this variation [25].

A number of studies in normal individuals have demonstrated associations between IL-10 secretion and microsatellite or SNP polymorphisms characterizing distinct IL-10 haplotypes [17, 20]. Other studies have suggested that some of these haplotypes are associated with inflammatory diseases [13, 21, 28], although there is no really convincing evidence that these polymorphisms are directly involved in influencing IL-10 production or disease susceptibility. However, they may be markers for other relevant mutations within the gene and it is conceivable that only certain IL10.R2 haplotypes include specific SNPs associated with increased IL-10 secretion. Equally, only certain IL10.R3 haplotypes may truly be under-represented and associated with reduced IL-10 secretion. Discrimination between the various IL10.R2 or IL10.R3 extended haplotypes is not possible in a case–control study. As current evidence regarding a possible association of IL-10 with RA is inconclusive it would be appropriate to conduct a within-family association study to define the effects of specific haplotypes.

In conclusion, this study did not confirm an association between IL10.R2 and RA. However, as the data are inconclusive, further large studies investigating IL-10 as a candidate gene are justified but will require large numbers to achieve adequate statistical power. Extended haplotyping of the IL-10 promoter region should help to define any disease-causing haplotypes and so improve the chances of identifying an association.

Acknowledgements

Kirsten MacKay was an Arthritis Research Campaign Clinical Research Fellow, Stephen Laval was funded by the Oliver Bird Fund of The Nuffield Foundation and Anita Milicic is sponsored by the Grenville Bequest. The financial support of the Arthritis Research Campaign and the Nuffield Foundation is greatly appreciated by the authors. We are grateful to Drs D. G. I. Scott, R. Sturrock, R. Madhok and P. Bacon for allowing us to study their patients with rheumatoid vasculitis.

References

1. Lipsky PE, van der Heijde DM, St Clair EW *et al.* Infliximab and methotrexate in the treatment of rheumatoid arthritis. N Engl J Med 2000;343:1594–602.
2. Arend WP, Guthridge CJ. Biological role of interleukin 1 receptor antagonist isoforms. Ann Rheum Dis 2000; 59(Suppl. 1):I60–4.
3. Weckmann AL, Alcocer-Varela J. Cytokine inhibitors in autoimmune disease. Semin Arthritis Rheum 1996; 26:539–57.
4. Keystone E, Wherry J, Grint P. IL-10 as a therapeutic strategy in the treatment of rheumatoid arthritis. Rheum Dis Clin North Am 1998;24:629–39.
5. Woods JM, Tokuhira M, Berry JC *et al.* Interleukin-4 adenoviral gene therapy reduces production of inflammatory cytokines and prostaglandin E2 by rheumatoid arthritis synovium *ex vivo*. J Investig Med 1999;47:285–92.
6. de Waal Malefyt R, Abrams J, Bennett B, Figdor CG, de Vries JE. Interleukin 10(IL-10) inhibits cytokine synthesis by human monocytes: an autoregulatory role of IL-10 produced by monocytes. J Exp Med 1991; 174:1209–20.
7. Burdin N, Peronne C, Banchereau J, Rousset F. Epstein–Barr virus transformation induces B lymphocytes to produce human interleukin 10. J Exp Med 1993; 177:295–304.
8. Rousset F, Garcia E, Defrance T *et al.* Interleukin 10 is a potent growth and differentiation factor for activated human B lymphocytes. Proc Natl Acad Sci USA 1992; 89:1890–3.

9. Katsikis PD, Chu CQ, Brennan FM, Maini RN, Feldmann M. Immunoregulatory role of interleukin 10 in rheumatoid arthritis. J Exp Med 1994;179:1517–27.
10. Westendorp RG, Langermans JA, Huizinga TW *et al.* Genetic influence on cytokine production and fatal meningococcal disease [published erratum appears in Lancet 1997;349:656] [see comments]. Lancet 1997; 349:170–3.
11. Kurtzhals JA, Adabayeri V, Goka BQ *et al.* Low plasma concentrations of interleukin 10 in severe malarial anaemia compared with cerebral and uncomplicated malaria [published errata appear in Lancet 1998;352:242 and 1999;353:848]. Lancet 1998;351:1768–72.
12. Eskdale J, Kube D, Tesch H, Gallagher G. Mapping of the human IL10 gene and further characterization of the 5′ flanking sequence. Immunogenetics 1997;46:120–8.
13. Crawley E, Kay R, Sillibourne J, Patel P, Hutchinson I, Woo P. Polymorphic haplotypes of the interleukin-10 5′ flanking region determine variable interleukin-10 transcription and are associated with particular phenotypes of juvenile rheumatoid arthritis. Arthritis Rheum 1999; 42:1101–8.
14. Huizinga TW, Keijsers V, Yanni G *et al.* Are differences in interleukin 10 production associated with joint damage? Rheumatology 2000;39:1180–8.
15. Eskdale J, Gallagher G. A polymorphic dinucleotide repeat in the human IL-10 promoter. Immunogenetics 1995;42:444–5.
16. Eskdale J, Kube D, Gallagher G. A second polymorphic dinucleotide repeat in the 5′ flanking region of the human IL10 gene. Immunogenetics 1996;45:82–3.
17. Eskdale J, Gallagher G, Verweij CL, Keijsers V, Westendorp RG, Huizinga TW. Interleukin 10 secretion in relation to human IL-10 locus haplotypes. Proc Natl Acad Sci USA 1998;95:9465–70.
18. Eskdale J, Keijsers V, Huizinga T, Gallagher G. Microsatellite alleles and single nucleotide polymorphisms (SNP) combine to form four major haplotype families at the human interleukin-10 (IL-10) locus. Genes Immun 1999;1:151–5.
19. D'Alfonso S, Rampi M, Rolando V, Giordano M, Momigliano-Richiardi P. New polymorphisms in the IL-10 promoter region. Genes Immun 2000;1:213–33.
20. Turner DM, Williams DM, Sankaran D, Lazarus M, Sinnott PJ, Hutchinson IV. An investigation of polymorphism in the interleukin-10 gene promoter. Eur J Immunogenet 1997;24:1–8.
21. Eskdale J, Wordsworth P, Bowman S, Field M, Gallagher G. Association between polymorphisms at the human IL-10 locus and systemic lupus erythematosus [published erratum appears in Tissue Antigens 1997;50:699]. Tissue Antigens 1997;49:635–9.
22. Eskdale J, McNicholl J, Wordsworth P *et al.* Interleukin-10 microsatellite polymorphisms and IL-10 locus alleles in rheumatoid arthritis susceptibility [letter]. Lancet 1998; 352:1282–3.
23. Spurdle A, Jenkins T. Y chromosome probe p49a detects complex PvuII haplotypes and many new TaqI haplotypes in southern African populations. Am J Hum Genet 1992; 50:107–25.
24. Olerup O, Zetterquist H. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 h: an alternative to serological DR typing in clinical practice including donor–recipient matching in cadaveric transplantation. Tissue Antigens 1992;39:225–35.
25. Rudwaleit M, Tikly M, Khamashta M *et al.* Inter-ethnic differences in the association of tumor necrosis factor promoter polymorphisms with systemic lupus erythematosus. J Rheumatol 1996;23:725–8.
26. Fei BY, Deng CS, Xia B, Zhu YQ, Crusius JB, Pena AS. Polymorphisms at the TNF locus in Chinese Han population. Hum Immunol 2002;63:71–5.
27. Gallagher G, Eskdale J, Oh HH, Richards SD, Campbell DA, Field M. Polymorphisms in the TNF gene cluster and MHC serotypes in the West of Scotland. Immunogenetics 1997;45:188–94.
28. Mozzato-Chamay N, Mahdi OS, Jallow O, Mabey DC, Bailey RL, Conway DJ. Polymorphisms in candidate genes and risk of scarring trachoma in a *Chlamydia trachomatis*-endemic population. J Infect Dis 2000;182:1545–8.